LOCAL EXCHANGEABILITY

By Trevor Campbell[*], Saifuddin Syed[*], Chiao-Yu Yang[†], Michael I. Jordan[†], and Tamara Broderick[‡]

University of British Columbia[*], *University of California, Berkeley*[†], *and Massachusetts Institute of Technology*[‡]

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 probabilistic models and efficient inference algorithms. In practice, however, this assumption is too strong an idealization; the distribution typically fails to be *exactly* invariant to permutations and de Finetti's representation theory does not apply. Thus there is the need for a distributional assumption that is both weak enough to hold in practice, and strong enough to guarantee a useful underlying representation. We introduce a relaxed notion of *local exchangeability*—where swapping data associated with nearby covariates causes a bounded change in the distribution. Next, we prove that locally exchangeable processes correspond to independent observations from an underlying measure-valued stochastic process, showing that de Finetti's theorem is robust to perturbation and providing further justification for the Bayesian modelling approach. We also provide an investigation of approximate sufficiency and sample continuity properties of locally exchangeable processes on the real line. The paper concludes with examples of popular statistical models that exhibit local exchangeability.

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. For example, if one were to randomly select a sequence of i.i.d. dates throughout a year and encode whether it was raining that day as a $\{0, 1\}$ variable, one would not expect the probability of seeing $1, 1, 0$ to differ from $0, 1, 1$ or $1, 0, 1$. The same intuition holds for a wide variety of sequential repeated experiments.

Keywords and phrases: exchangeability, local, representation, de Finetti

Despite its apparent innocence, however, 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 measure G on $\mathcal{X}$ such that

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

where G^{∞} is the countable infinite product measure constructed from G. This result not only exposes the fact that the sequence X has a conditional independence structure useful for computationally efficient inference; it also provides a strong justification for the Bayesian approach to modeling when dealing with exchangeable data (Jordan, 2010). See Aldous (1985) for an excellent introduction to exchangeability and related topics. In the weather recording example, this result guarantees that there is a latent, random probability $p \in [0, 1]$ with which it rains independently on each selected day. We are therefore justified in placing a prior on p and taking a Bayesian approach to learning how rainy it is at that location.

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), correspondences between probabilistic invariances and conditional latent structure (known as *representation theorems*) have been studied extensively. For example, it has been shown that a *finite* exchangeable sequence corresponds to a mixture of urn models, and that if the sequence is also extendable it is close to a mixture of i.i.d. sequences (Diaconis, 1977; Diaconis and Freedman, 1980a). Further notions of exchangeability and corresponding latent conditional structure now exist for a wide variety of probabilistic models, such as arrays (Aldous, 1981; Hoover, 1979), Markov processes (Diaconis and Freedman, 1980b), networks (Caron and Fox, 2017; Veitch and Roy, 2015; Borgs et al., 2018; Crane and Dempsey, 2016a; Cai, Campbell and Broderick, 2016; Janson, 2017), combinatorial structures (Kingman, 1978; Pitman, 1995; Broderick, Pitman and Jordan, 2013; Campbell, Cai and Broderick, 2018; Crane and Dempsey, 2016b), random measures (Kallenberg, 1990), and more (Diaconis, 1988; Kallenberg, 2005; Orbanz and Roy, 2015).

However, although exchangeability may be a useful idealization for modeling and computational purposes, it rarely holds for real data sequences; typically there are differences in the sequence of observations that preclude an honest belief in exchangeability. In the weather recording example, for instance, the sequence was exchangeable only because we selected dates i.i.d. and discarded the (very relevant) date covariate information. But despite a lack of exact exchangeability, we do still expect observations with similar covariates to have a similar law. de Finetti (1938, English translation, p. 198) actually raises this point, noting that "we must take up the case where we still encounter 'analogies' among the events under consideration, but without their attaining the limiting case of exchangeability." And there are many examples of popular statistical models whose observations intuitively exhibit such similarities without satisfying exact exchangeability, such as Gaussian process regression (Rasmussen and Williams, 2006), dynamic topic models with slowly varying latent topics (Blei and Lafferty, 2006; Wang, Blei and Heckerman, 2008), as well as covariant-dependent clustering (MacEachern, 1999, 2000) and feature allocation (Ren et al., 2011).

One option is *partial exchangeability* (de Finetti, 1938; Lauritzen, 1974; Diaconis and Freedman, 1978; Camerlenghi et al., 2019): 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 equal 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 measures $(G_t)_{t\in\mathcal{T}}$,

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

The random 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. Thus, 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. Next, 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 is the first representation theorem 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 locally exchangeable data. The paper then provides a study of the special case of local exchangeability on the real line, drawing connections with stationarity and path continuity of the underlying measure-valued process. We also investigate the analogue of the sufficiency of the empirical measure in classical sequence exchangeability, showing that inexact covariate information due to discretization (or "binning") has information content that degrades smoothly with the discretization coarseness. The paper concludes with examples of statistical models that exhibit local exchangeability, including Gaussian processes with noisy observations (Rasmussen and Williams, 2006), dependent Dirichlet process mixture models (MacEachern, 1999, 2000), kernel beta process latent feature models (Ren et al., 2011), and dynamic topic models (Blei and Lafferty, 2006; Wang, Blei and Heckerman, 2008).

2. Local exchangeability. Let $X = (X_t)_{t\in\mathcal{T}}$ be a stochastic process on a set $\mathcal{T}$ taking values in a standard Borel space $(\mathcal{X}, \Sigma)$. We endow the covariate space $\mathcal{T}$ with a pseudometric $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ (distances between distinct points may be 0) to capture a notion of locality of observations. Typically the pseudometric will follow naturally from the problem setting. For example, suppose X are the noisy observations from a Gaussian process regression model with input space $\mathbb{R}$; here the Euclidean metric is the natural choice to encode locality. But if obtaining multiple observations at the same spatial point is possible, we must choose the stochastic process index set $\mathcal{T}$ to allow for this. An appropriate choice is $\mathcal{T} = \mathbb{R} \times \mathbb{N}$, where the first component denotes the location of the observation, and the second denotes the index of the observation at that location; e.g., $(0.4, 3)$ would denote the 3^{rd} observation at location 0.4. In this case, the pseudometric is $d(t, t') = |x - x'|$ for $t = (x, n)$, $t' = (x', n')$, $x, x' \in \mathbb{R}$, $n, n' \in \mathbb{N}$—such that observations at the same spatial point are thought to be exactly exchangeable, while observations taken at points of increasing distance from one another are thought to be progressively less exchangeable. On the other hand, if having multiple observations at one location is not possible, setting $\mathcal{T} = \mathbb{R}$ and d to the Euclidean distance would suffice.

We will formalize local exchangeability based on the finite dimensional projections of the stochastic process X. For any finite subset $T \subset \mathcal{T}$, define the restriction $X_T = (X_t)_{t\in T}$, and for any bijection $\pi : T \to T$, denote the corresponding permuted random element by $X_{\pi T}$. 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 elements of T and πT are small.

DEFINITION 1. The process X is f-*locally exchangeable* if there exists a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\lim_{x\to 0} f(x) = f(0) = 0$ such that for any finite subset $T \subset \mathcal{T}$ and bijection $\pi : T \to T$,

$$d_{\mathrm{TV}}(X_T, X_{\pi T}) \leq \sum_{t\in T} f\left(d(t, \pi(t))\right). \tag{2}$$

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

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

The choice of total variation distance (as opposed to other metrics and divergences, see e.g. (Gibbs and Su, 2002)) is motivated by its generality and strength. In particular, it imposes no additional structure beyond that of a measurable space on $(\mathcal{X}, \Sigma)$, and it is strong enough to guarantee a de Finetti-like result below in Theorem 4.

To determine how distances $d(t, \pi(t))$ between covariates translate into total variation between finite dimensional distributions, Definition 1 uses a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ for which $f(x) \to 0$ as $x \to 0$; the rate of decay of f controls the "smoothness" of local exchangeability. By choosing different combinations of function f and pseudometric d, Definition 1 generalizes both partial and exact exchangeability. For example, the function $f(x) = x$ paired with the discrete pseudometric ($d(t, t') = 0$ if t and t' are in an equivalence class, and otherwise $d(t, t') = 1$) yields partial exchangeability. The same function paired with the zero metric (where $d(t, t') = 0$) yields classical exchangeability. The fact that d may be a pseudometric, as a opposed to a metric, is critical for these natural generalizations to hold.

Note that Definition 1 has two key advantages compared with other possible methods of controlling total variation between finite marginals of a locally exchangeable process. First, it extends to the case of countable covariate sets and injection mappings via Proposition 2.

Proposition 2. *If X is f-locally exchangeable, then for any countable subset $T_\infty \subset \mathcal{T}$ and injection $\pi : \mathcal{T} \to \mathcal{T}$,*

$$d_{\mathrm{TV}}(X_{T_\infty}, X_{\pi T_\infty}) \leq \sum_{t\in T_\infty} 2f\left(d(t, \pi(t))\right).$$

Proof. Choose some ordering $T_\infty = (t_1, t_2, \dots)$ and any $A_\infty \in \Sigma^\infty$. By the continuity of measures,

$$|\mathbb{P}(X_{T_\infty} \in A_\infty) - \mathbb{P}(X_{\pi T_\infty} \in A_\infty)| = \lim_{n\to\infty} |\mathbb{P}(X_{t_{1:n}} \in A_n) - \mathbb{P}(X_{t'_{1:n}} \in A_n)|,$$

where A_n is the projection of A_∞ onto the first n indices $t_1, \ldots, t_n$, and $t'_i := \pi t_i$ for brevity. If $T := \{t_1, \ldots, t_n\}$, $T' := \{t'_1, \ldots, t'_n\}$, $m := |T \cup T'|$, $t_{n+1}, \ldots, t_m$ is an ordering of $T' \setminus T$, and $t'_{n+1}, \ldots, t'_m$ is an ordering of $T \setminus T'$, we have that

$$\mathbb{P}(X_{t_{1:n}} \in A_n) = \mathbb{P}\left(X_{t_{1:m}} \in A_n \times \mathcal{X}^{m-n}\right)$$
$$\mathbb{P}(X_{t'_{1:n}} \in A_n) = \mathbb{P}\left(X_{t'_{1:m}} \in A_n \times \mathcal{X}^{m-n}\right).$$

Thus by the triangle inequality and local exchangeability with the permutation that swaps t and $\pi(t)$ for $t \in T$,

$$|\mathbb{P}(X_{t_{1:n}} \in A_n) - \mathbb{P}(X_{t'_{1:n}} \in A_n)| \leq \sum_{t \in T \cup T'} f\left(d(t, \pi(t))\right) \leq \sum_{t \in T} 2f\left(d(t, \pi(t))\right).$$

Thus the limit is also bounded by this quantity, and the result follows. □

Second, Definition 1 has an easy-to-verify sufficient condition given by Proposition 3. This is useful because it can be cumbersome to determine whether a process is f-locally exchangeable, not to mention picking a suitable function f. Proposition 3 shows that for a general class of processes, f can be determined by the relationship between the chosen pseudometric d and the *canonical pseudometric* $d_c : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$ of the process. In particular, if the values of X are conditionally independent given some underlying distributions indexed by $\mathcal{T}$, then f may be selected based on how quickly the mean total variation between those distributions varies across $\mathcal{T}$. This technique will be used extensively in Section 3 for the analysis of popular statistical models.

PROPOSITION 3. *Suppose there exists a σ-algebra $\mathcal{G}$ such that the values of the process X at any finite subset of covariates T are conditionally independent given $\mathcal{G}$, and let G_t be the conditional distribution of X_t given $\mathcal{G}$. Then X is f-locally exchangeable if*

$$\forall t, t' \in \mathcal{T} \qquad d_c(t, t') := \mathbb{E}\left[d_{\mathrm{TV}}(G_t, G_{t'})\right] \leq f(d(t, t')).$$

REMARK. Since $(\mathcal{X}, \Sigma)$ is a standard Borel space, the random measure G_t is guaranteed to exist (Kallenberg, 2002, Theorem 6.3). Further, $d_{\mathrm{TV}}(G_t, G_{t'})$ is guaranteed to be measurable: by separability of $\mathcal{X}$, the open sets have a countable base, which generates a countable algebra, which then generates Σ, and by Halmos (1978, Theorem D, p. 56) it suffices to take the supremum over sets in the countable algebra.

PROOF. Defining $G_T := \prod_{t\in T} G_t$ to be the conditional distribution of X_T given $\mathcal{G}$,

$$\begin{aligned} d_{\mathrm{TV}}(X_T, X_{\pi T}) &= \sup_A |\mathbb{P}(X_T \in A) - \mathbb{P}(X_{\pi T} \in A)| \\ &= \sup_A |\mathbb{E}\left[G_T(A) - G_{\pi T}(A)\right]| \\ &\le \mathbb{E}\left[d_{\mathrm{TV}}(G_T, G_{\pi T})\right]. \end{aligned}$$

By Lemma 8, $d_{\mathrm{TV}}(X_T, X_{\pi T}) \le \sum_{t\in T} \mathbb{E}\left[d_{\mathrm{TV}}(G_t, G_{\pi(t)})\right] \le \sum_{t\in T} f(d(t, \pi(t)))$. $\square$

It remains to show whether Definition 1 is a useful assumption in practical modeling and computation, and whether it provides justification for a Bayesian approach. One way of achieving this would be to obtain a de Finetti-like representation of a form similar to Eq. (1). Theorem 4 shows that such a representation indeed does exist under fairly general conditions: as long as the covariate space $\mathcal{T}$ is "nice" enough, for any f-locally exchangeable process X, (a) there is a unique probability measure-valued process G that renders X conditionally independent, and (b) G satisfies a continuity property with the same "smoothness" as the observed process. For the precise statement of the result in Theorem 4, recall that a *modification* of a stochastic process G on $\mathcal{T}$ is any other process G' on $\mathcal{T}$ such that $\forall t \in \mathcal{T}$, $\mathbb{P}\left(G_t = G'_t = 1\right)$.

THEOREM 4. *Suppose $\mathcal{T}$ is separable with no isolated points under the pseudometric d. Then the process X is f-locally exchangeable iff 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 any permutation $\pi : T \to T$,*

$$\mathbb{P}\left(X_T \in \cdot \mid G\right) \overset{a.s.}{=} G_T \qquad G_T := \prod_{t\in T} G_t$$

and

$$\sup_A \mathbb{E}\left|G_T(A) - G_{\pi T}(A)\right| \le \sum_{t\in T} f(d(t, \pi(t))). \tag{3}$$

REMARK. The proof of Theorem 4 shows that G is measurable with respect to the tail σ-algebra of any countable dense subset of $\mathcal{T}$.

PROOF. The reverse direction follows from the triangle inequality:

$$\sup_A |\mathbb{P}(X_T \in A) - \mathbb{P}(X_{\pi T} \in A)| = \sup_A |\mathbb{E}\left[G_T(A) - G_{\pi T}(A)\right]|$$

$$\leq \sup_A \mathbb{E}\,|G_T(A) - G_{\pi T}(A)|$$
$$\leq \sum_{t \in T} f(d(t, \pi(t))).$$

For the forward direction, suppose X is locally exchangeable, let $(t_n)_{n=1}^{\infty}$ be a countable dense subset of $\mathcal{T}$, and let $\mathcal{F}$ be the tail σ-algebra of $(X_{t_n})_{n=1}^{\infty}$. We will show that for any two covariates $r, s \in \mathcal{T}$, $r \neq s$, $r, s \notin \{t_n\}_{n=1}^{\infty}$, X_r and X_s are conditionally independent given $\mathcal{F}$. The argument extends via standard methods to r, s that may be elements of $\{t_n\}_{n=1}^{\infty}$, and then to any finite subset of $\mathcal{T}$.

Since $f(x) \to 0$ as $x \to 0$ and the sequence $(t_n)_{n=1}^{\infty}$ is dense in $\mathcal{T}$, there exists a subsequence $i_1 < i_2 < \dots$ of indices such that $(t_{i_n})_{n=1}^{\infty}$ converges to s, and for all $N \in \mathbb{N}$, $i_N > N$ and $f(d(s, t_{i_N})) + \sum_{n=N}^{\infty} f(d(t_{i_n}, t_{i_{n+1}})) < 1/(2N)$. Let π_N be the mapping that takes $s \to t_{i_N}$, $t_{i_n} \to t_{i_{n+1}}$ for all $n \geq N$, and leaves all other $t \in \mathcal{T}$ fixed. Then denote $Y_N = (X_s, X_{t_N}, X_{t_{N+1}}, \dots)$, and let Z_N be the sequence with covariates mapped under π_N. By reverse martingale convergence, for any bounded $\phi : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}\left[\phi(X_r) \,|\, Y_N\right] \overset{a.s.}{\to} \mathbb{E}\left[\phi(X_r) \,|\, X_s, \mathcal{F}\right] \quad \text{and} \quad \mathbb{E}\left[\phi(X_r) \,|\, Z_N\right] \overset{a.s.}{\to} \mathbb{E}\left[\phi(X_r) \,|\, \mathcal{F}\right]$$

as $N \to \infty$. Next, by local exchangeability and Proposition 2,

$$d_{\mathrm{TV}}((X_r, Y_N), (X_r, Z_N)) < \frac{1}{N},$$

and by Lemma 9(2), we have that the Wasserstein distance between $\mathbb{E}\left[\phi(X_r) \,|\, Y_N\right]$ and $\mathbb{E}\left[\phi(X_r) \,|\, Z_N\right]$ converges to 0 as $N \to \infty$. Thus combined with the previous reverse martingale result,

$$\mathbb{E}\left[\phi(X_r) \,|\, X_s, \mathcal{F}\right] \overset{d}{=} \mathbb{E}\left[\phi(X_r) \,|\, \mathcal{F}\right].$$

By Aldous (1985, Lemma 3.4),

$$\mathbb{E}\left[\phi(X_r) \,|\, X_s, \mathcal{F}\right] \overset{a.s.}{=} \mathbb{E}\left[\phi(X_r) \,|\, \mathcal{F}\right],$$

and thus X_r and X_s are conditionally independent given $\mathcal{F}$. As mentioned earlier this argument extends easily to any finite subset T of covariates, by considering subsequences of $(t_n)_{n=1}^{\infty}$ converging to each $t \in T$. Since X takes values in a standard Borel space, there is a random measure G_t for each $t \in \mathcal{T}$ for which $G_t(A) \overset{a.s.}{=} \mathbb{E}\left[\mathbb{1}[X_t \in A] \,|\, \mathcal{F}\right]$ (e.g. Kallenberg, 2002, Theorem 6.3). The collection of these random measures forms the desired stochastic process $G = (G_t)_{t \in \mathcal{T}}$.

Next, we develop the smoothness property of G. By both reverse and forward martingale convergence, we have that

$$\begin{aligned}&\sup_A \mathbb{E}\left|G_T(A) - G_{\pi T}(A)\right| \\ &\qquad = \sup_A \mathbb{E}\left|\lim_{n\to\infty}\lim_{m\to\infty} \mathbb{E}\left[\mathbb{1}\left[X_T \in A\right] - \mathbb{1}\left[X_{\pi T} \in A\right] \mid X_{t_{n:n+m}}\right]\right|,\end{aligned}$$

Using dominated convergence to move the limits out of the expectation, local exchangeability to bound the total variation between $(X_T, X_{t_{n:n+m}})$ and $(X_{\pi T}, X_{t_{n:n+m}})$, and Lemma 9(1),

$$\sup_A \mathbb{E}\left|G_T(A) - G_{\pi T}(A)\right| \leq \sum_{t\in T} f(d(t, \pi(t))). \tag{4}$$

Finally, we show that G is unique up to modification, and that it doesn't depend on our choice of the dense subset $(t_n)_{n=1}^\infty$. Consider any $A \in \Sigma$ and any sequence $(t'_n)_{n=1}^\infty$ converging to $s \in \mathcal{T}$ such that $f(d(t'_n, s)) \leq 2^{-n}$ for each $n \in \mathbb{N}$. Define $S_{s,N} = \frac{1}{N}\sum_{n=1}^N \mathbb{1}[X_{t'_n} \in A]$. Then

$$\begin{aligned}&\mathbb{P}\left(\left|S_{s,N} - G_s(A)\right| > \epsilon\right) \\ &\quad = \mathbb{E}\left[\mathbb{P}\left(\left|S_{s,N} - \frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) + \frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) - G_s(A)\right| > \epsilon \,|\, \mathcal{F}\right)\right] \\ &\quad \leq \mathbb{E}\left[\mathbb{P}\left(\left|S_{s,N} - \frac{1}{N}\sum_{n=1}^N G_{t'_n}(A)\right| + \left|\frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) - G_s(A)\right| > \epsilon \,|\, \mathcal{F}\right)\right].\end{aligned}$$

Noting that the right term is $\mathcal{F}$-measurable and applying Hoeffding's inequality to the left,

$$\mathbb{P}\left(\left|S_{s,N} - G_s(A)\right| > \epsilon\right) \leq \mathbb{E}\left[2e^{-2N\left(\max\left\{0, \epsilon - \left|\frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) - G_s(A)\right|\right\}\right)^2}\right].$$

Splitting the above expectation across two events—one where the measures satisfy $\left|\frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) - G_s(A)\right| > \epsilon/2$ and the other its complement—yields

$$\mathbb{P}\left(\left|S_{s,N} - G_s(A)\right| > \epsilon\right) \leq \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^N G_{t'_n}(A) - G_s(A)\right| > \frac{\epsilon}{2}\right) + 2e^{-N\epsilon^2/2}.$$

Applying Markov's inequality, the triangle inequality, and Eq. (4),

$$\mathbb{P}\left(\left|S_{s,N} - G_s(A)\right| > \epsilon\right) \leq \frac{2}{\epsilon N}\sum_{n=1}^N \mathbb{E}\left|G_{t'_n}(A) - G_s(A)\right| + 2e^{-N\epsilon^2/2}$$

$$\leq \frac{2}{\epsilon N}\sum_{n=1}^{N} 2^{-n} + 2e^{-N\epsilon^2/2} \to 0 \quad \text{as} \quad N \to \infty.$$

Thus, $S_{s,N} \overset{p}{\to} G_s(A)$. By Kallenberg (2002, Lemma 4.2), there exists a subsequence $S_{s,N_i} \overset{a.s.}{\to} G_s(A)$ as $i \to \infty$. One can repeat this procedure for the countable algebra that generates Σ, and then by Carathéodory's extension theorem (Kallenberg, 2002, Theorem 2.5), G_s is uniquely determined. The extension of this technique to any finite subset of covariates $T \subset \mathcal{T}$ is straightforward, implying that $(G_t)_{t\in\mathcal{T}}$ is uniquely determined up to modification. □

2.1. *Regularity.* The smoothness property of G in Eq. (3) 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 (Lin and Fisher, 2010; Chen et al., 2013) have components that are created and destroyed over time, causing discrete jumps in the mixing measure. But if the jumps happen at diffuse random times, the probability of a jump occurring between two times decreases as the difference in time decreases, rendering the marginal distribution of the observations locally exchangeable. In general, as long as any discontinuities have a diffuse random location in $\mathcal{T}$, the process may still be locally exchangeable.

That being said, it is worth examining whether different guarantees on properties of the underlying measure process G result as a consequence of different settings of the function f. Theorem 5 answers this question in the affirmative for processes on $\mathcal{T} = \mathbb{R}$; in particular, the faster the decay of $f(x)$ as $x \to 0$, 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 (Pothoff, 2009, Theorems 2.8, 2.9, 4.5).

THEOREM 5. *Let X be an f-locally exchangeable stochastic process on $\mathcal{T} = \mathbb{R}$, d be the Euclidean metric $d(t,t') = |t - t'|$, and $f(x) = O(x^{1+\gamma})$ as $x \to 0$ for $\gamma \geq 0$. Then Theorem 4 holds, and:*

1. *($\gamma > 1$): G is a constant process, and X is exchangeable.*
2. *($0 < \gamma \leq 1$): G is locally α-Hölder continuous in the weak topology for any $\alpha \in (0,\gamma)$ and weak-sense stationary, and X is stationary.*

3. ($\gamma = 0$): G may have no continuous modification.

REMARK. A rough converse of the first point holds: X exchangeable implies constant G and $f(x) = 0$ is trivially $O(x^{1+\gamma})$ for $\gamma > 1$. But a similar claim for the second point is not true in general: X stationary and f-locally exchangeable does not necessarily imply that $f(x) = O(x^{1+\gamma})$ for $0 < \gamma \leq 1$. For a counterexample, consider the square wave shifted by a uniform random variable, i.e., the process $X_t = \operatorname{sgn}\sin(2\pi(t-U))$ for $U \sim \mathsf{Unif}[0,1]$. Here X_t is stationary and f-locally exchangeable with $f(x) = x$ by Proposition 3, but $x \neq O(x^{1+\gamma})$ for any $\gamma > 0$ as $x \to 0$.

PROOF. Using the result of Proposition 2, for any finite set of covariates $T \subset \mathbb{R}$, measurable set A, and $t, \Delta \in \mathbb{R}$,

$$|\mathbb{P}(X_T \in A) - \mathbb{P}(X_{T+\Delta} \in A)| \leq 2|T| f(\Delta) = O(\Delta^{1+\gamma}), \quad \Delta \to 0,$$

and

$$\mathbb{E}|G_t(A) - G_{t+\Delta}(A)| \leq 2f(\Delta) = O(\Delta^{1+\gamma}), \quad \Delta \to 0, \tag{5}$$

where $T + \Delta$ denotes the translation of all covariates in T by Δ. The Kolmogorov continuity theorem (Kallenberg, 2002, Theorem 3.23) applied to Eq. (5) implies that for all $\alpha \in (0, \gamma)$ and $A \in \Sigma$, $(G_t(A))_{t\in\mathbb{R}}$ is locally α-Hölder continuous. Hence if $\gamma > 1$, G is constant, and thus X is exchangeable. Similarly if $\gamma \in (0, 1]$, G is locally α-Hölder continuous in the weak topology. Furthermore,

$$\lim_{\Delta\to 0} \frac{|\mathbb{P}(X_T \in A) - \mathbb{P}(X_{T+\Delta} \in A)|}{\Delta} \leq C \cdot \lim_{\Delta \to 0} \Delta^\gamma = 0,$$

showing that X is stationary. Next, since X is stationary, for any $t, t' \in \mathbb{R}$ and $A \in \Sigma$, the mean of $G_t(A)$ satisfies

$$\mathbb{E}\left[G_t(A)\right] = \mathbb{P}(X_t \in A) = \mathbb{P}(X_{t+\Delta} \in A) = \mathbb{E}\left[G_{t+\Delta}(A)\right].$$

Similarly, the autocovariance satisfies

$$\begin{aligned}
&\mathbb{E}\left[(G_t(A) - \mathbb{E}G_t(A))(G_{t+\Delta}(A) - \mathbb{E}G_{t+\Delta}(A))\right] \\
&\qquad = \mathbb{P}(X_t \in A, X_{t+\Delta} \in A) - \mathbb{P}(X_t \in A)\mathbb{P}(X_{t+\Delta} \in A) \\
&\qquad = \mathbb{P}(X_0 \in A, X_\Delta \in A) - \mathbb{P}(X_0 \in A)\mathbb{P}(X_\Delta \in A) \\
&\qquad = \mathbb{E}\left[(G_0(A) - \mathbb{E}G_0(A))(G_\Delta(A) - \mathbb{E}G_\Delta(A))\right].
\end{aligned}$$

Hence $(G_t(A))_{t\in\mathbb{R}}$ is weak-sense stationary. Finally, note that the process $X_t = \mathbb{1}(t \geq U)$ for $U \in \mathsf{Unif}[0,1]$ is locally exchangeable with $f(x) = x$ by Proposition 3, but its underlying random measure specified by $G_t = \mathbb{1}(t < U)\delta_{\{0\}} + \mathbb{1}(t \geq U)\delta_{\{1\}}$ where δ_x is the Dirac measure at x has no sample-continuous modification. $\square$

2.2. *Approximate sufficiency.* In the classical setting of exchangeable sequences $X_1, X_2, \dots,$ de Finetti's theorem (1) not only guarantees the existence of a mixing measure G, but also shows that the empirical measure $M_N = \sum_{n=1}^N \delta_{X_n}$ is a sufficient statistic for G. In other words, for all bounded measurable functions $h : \mathcal{X}^N \to \mathbb{R}$,

$$\mathbb{E}\left[h(X_1, \dots, X_N) | M_N, G\right] = \mathbb{E}\left[h(X_1, \dots, X_N) | M_N\right].$$

Thus the conditional distribution of G given the observed data is the same whether we have knowledge of M_N or $(X_1, \dots, X_N)$ 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 don't often get to observe the true covariate values $\{t_1, \dots, t_N\} = T \subset \mathcal{T}$, but rather 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 quantity analogous to the empirical measure is approximately sufficient for G, and 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 N_k 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. We will encode our lack of precise knowledge of T 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 $M_k = \sum_{t \in T_k} \delta_{X_t}$, $M_T := (M_1, M_2, \dots)$, and let S_T denote the group 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 $|S_T| = \prod_{k=1}^\infty N_k! < \infty$ since there are only finitely many observations in total. Unlike classical exchangeability, M_T is in general not a sufficient statistic for G; but Theorem 6 guarantees that M_T is approximately sufficient, with error that depends on $(\mu_k)_{k=1}^\infty$.

THEOREM 6. *If X is f-locally exchangeable, and $h : \mathcal{X}^T \to \mathbb{R}$ is a bounded measurable function,*

$$\mathbb{E}\left|\mathbb{E}\left[h(X_T) \mid M_T, G\right] - \mathbb{E}\left[h(X_T) \mid M_T\right]\right| \le 4\|h\|_\infty \mathbb{E}\left[\sum_{t \in T} f(d(t, \pi(t)))\right], \tag{6}$$

where $\pi \sim \mathsf{Unif}\{S_T\}$ 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.

PROOF. Let $t_1, t_2, \ldots$ be a countable dense subset of $\mathcal{T}$ not containing T, and $Y_N = \left(X_{t_N}, X_{t_{N+1}}, \ldots\right)$. Reverse martingale convergence implies that

$$\mathbb{E}\left[h(X_T)|Y_N, M_T\right] \overset{a.s.}{\to} \mathbb{E}\left[h(X_T)|\mathcal{F}, M_T\right] \overset{a.s.}{=} \mathbb{E}\left[h(X_T)|G, M_T\right] \quad N \to \infty, \tag{7}$$

where $\mathcal{F}$ is the tail σ-algebra of $\{X_{t_i}\}_{i=1}^{\infty}$. Defining $g(X_T) = \frac{1}{|S_T|}\sum_{\pi \in S_T} h(X_{\pi T})$, we have that $g(X_T)$ is invariant to S_T and thus $g(X_T)$ is $\sigma(M_T, Y_N)$-measurable. Therefore

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}[h(X_T)|M_T, Y_N] - g(X_T)\right| \\
&= \mathbb{E}\left[\frac{1}{|S_T|}\left|\sum_{\pi \in S_T} \mathbb{E}[h(X_T) - h(X_{\pi T})|M_T, Y_N]\right|\right] \\
&\le \mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi \in S_T} \left|\mathbb{E}\left[h(X_T) - h(X_{\pi T})|M_T, Y_N\right]\right|\right].
\end{aligned} \tag{8}$$

By Lemma 9(1) and Proposition 2,

$$\begin{aligned}
&\le 2\|h\|_\infty \mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi \in S_T} d_{\mathrm{TV}}(X_T, X_{\pi T})\right] \\
&\le 2\|h\|_\infty \mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi \in S_T}\sum_{t \in T} f(d(t, \pi(t)))\right].
\end{aligned}$$

Taking the limit as $N \to \infty$, moving it into the expectation in Eq. (8) via dominated convergence, and using the limit from Eq. (7) yields

$$\mathbb{E}\left|\mathbb{E}[h(X_T)|M_T, G] - g(X_T)\right| \le 2\|h\|_\infty \mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi \in S_T}\sum_{t \in T} f(d(t, \pi(t)))\right].$$

Identical reasoning to the above also shows that

$$\mathbb{E}\left|\mathbb{E}[h(X_T)|M_T] - g(X_T)\right| \le 2\|h\|_\infty \mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi \in S_T}\sum_{t \in T} f(d(t, \pi(t)))\right].$$

Finally we add and subtract $g(X_T)$ in left hand side of Eq. (6), apply the triangle inequality with the above bounds, and note that the sum over π is the expectation over a uniformly random permutation to obtain the result. $\square$

If X is exchangeable within each bin $\mathcal{T}_k$, Theorem 6 states that M_T is exactly sufficient for G, as desired. Further, the deviance from sufficiency is controlled by the deviance from exchangeability within each bin. In particular,

$$\mathbb{E}\left[\frac{1}{|S_T|}\sum_{\pi\in S_T}\sum_{t\in T} f(d(t,\pi(t)))\right] \leq \sum_{k=1}^{\infty}|T_k| f(\operatorname{diam}\mathcal{T}_k) \tag{9}$$

$$\leq |T| f\left(\sup_k\{\operatorname{diam}\mathcal{T}_k\}\right), \tag{10}$$

where $\operatorname{diam}\mathcal{T}_k := \sup_{t,t'\in\mathcal{T}_k} d(t,t')$ and we extend f via $f(\infty) := \lim_{x\to\infty} f(x)$ if necessary. Both upper bounds are independent of μ_k; if we are unwilling to express our uncertainty in the binned covariates as a distribution, the bounds in Eqs. (9) and (10) still provide approximate sufficiency of M_T.

3. Examples of locally exchangeable processes. In this section, we detail some examples of popular statistical models that satisfy local exchangeability (but not exchangeability). While the main definition itself in Eq. (2) is often not easy to verify directly—due to the requirement that the condition holds for any finite set of covariates and any permutation—the sufficient condition in Proposition 3 only requires checking the latent random measure at pairs of covariates. Although doing so requires knowing in advance that such a latent random measure exists as well as its particular form, for many processes we do have this information, as detailed in the following.

3.1. *Gaussian processes.* For the first example of local exchangeability, we return to the Gaussian process setting introduced at the beginning of Section 2. Suppose the kernel function is $k(|x-x'|)$, $k:\mathbb{R}_+\to\mathbb{R}_+$ satisfying $k(y)\geq k(0)-ay^\gamma$ for some $a,\gamma\geq 0$. This is the case, for example, with the popular radial basis function $k(y)=e^{-y^2/2\ell^2}$ for $\gamma=2$ and $a=\frac{1}{2\ell^2}$. Then for a finite subset of covariates $T\subset\mathbb{R}\times\mathbb{N}$, the marginal distribution of the Gaussian process X is $X_T\sim\mathcal{N}(0,\sigma^2 I+K)$, where K is the $|T|\times|T|$ kernel matrix for subset T populated using the kernel function k. Comparing X_T and $X_{\pi T}$ directly is not trivial; however, this task can be simplified by rewriting the marginal distribution of X_T in the form $Y_T\sim\mathcal{N}(0,K)$ and $X_T\,|\,Y_T\sim\mathcal{N}(Y_T,\sigma^2 I)$ with X_T conditionally mutually independent given Y_T. In this setting, the canonical pseudometric is

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

Note that the total variation between two normal distributions may be expressed using the standard normal CDF Φ,

$$d_{\mathrm{TV}}(\mathcal{N}(Y_t,\sigma^2),\mathcal{N}(Y_{t'},\sigma^2)) = \Phi\left(\frac{|Y_t - Y_{t'}|}{2\sigma}\right) - \Phi\left(-\frac{|Y_t - Y_{t'}|}{2\sigma}\right).$$

Since the Lipschitz constant of the standard normal CDF Φ is $1/\sqrt{2\pi}$,

$$d_{\mathrm{TV}}(\mathcal{N}(Y_t,\sigma^2),\mathcal{N}(Y_{t'},\sigma^2)) \le \frac{|Y_t - Y_{t'}|}{\sqrt{2\pi\sigma^2}}. \tag{11}$$

Finally, Jensen's inequality applied to the expectation of Eq. (11) and the kernel lower bound together yield

$$d_c(t,t') \le \sqrt{\frac{2}{\pi\sigma^2}\left\{k(0) - k(d(t,t'))\right\}} \le \sqrt{\frac{2a}{\pi\sigma^2}}\, d(t,t')^{\frac{\gamma}{2}}.$$

Proposition 3 implies that X is locally exchangeable with $f(x) = \sqrt{2a/\pi\sigma^2}\, x^{\frac{\gamma}{2}}$.

3.2. *Dependent Dirichlet processes.* The next example of local exchangeability arises from Bayesian nonparametric mixture modelling. In a typical mixture model setting, we have observations generated via

$$X_n \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{\infty} w_k F(\cdot;\theta_k), \qquad\qquad n \in \mathbb{N},$$

where $(w_k)_{k=1}^{\infty}$ are the mixture weights satisfying $w_k \ge 0$, $\sum_k w_k = 1$, $(\theta_k)_{k=1}^{\infty}$ are the component parameters, $F(\cdot;\theta)$ is the mixture component likelihood, and $(X_n)_{n=1}^{\infty}$ are the observations. A popular nonparametric prior for the weights and component parameters is the Dirichlet process (Ferguson, 1973), defined 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. When the observations come with additional covariate information, the *dependent* Dirichlet process mixture model MacEachern (1999, 2000) may be used to capture similarities between related mixture population data. Here, observations are generated via

$$X_{x,n} \overset{\text{indep}}{\sim} \sum_{k=1}^{\infty} w_{x,k} F(\cdot;\theta_{x,k}), \qquad\qquad n \in \mathbb{N},\ x \in \mathbb{R},$$

where the component parameters $\theta_{x,k}$ and stick variables $v_{x,k}$ are now i.i.d. stochastic processes on $\mathbb{R}$, and $w_{x,k} = v_{x,k}\prod_{i=1}^{k-1}(1-v_{x,i})$. The marginal distributions of $\theta_{x,k}$ and $v_{x,k}$ at $x \in \mathbb{R}$ is H and $\mathsf{Beta}(1,\alpha)$, respectively. Thus, the dependent Dirichlet process is marginally a Dirichlet process for each covariate value, but can exhibit a wide range of dependencies across covariates. In this setting, we have $\mathcal{T} = \mathbb{R}\times\mathbb{N}$ and $d((x,n),(x',n')) = |x-x'|$, and the canonical pseudometric is

$$\begin{aligned} d_c(t,t') &= \mathbb{E}\left[d_{\mathrm{TV}}\left(\sum_{k=1}^{\infty} w_{x,k}F(\cdot;\theta_{x,k}), \sum_{k=1}^{\infty} w_{x',k}F(\cdot;\theta_{x',k})\right)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\int\left|\sum_{k=1}^{\infty} w_{x,k}F(y;\theta_{x,k}) - \sum_{k=1}^{\infty} w_{x',k}F(y;\theta_{x',k})\right|\mathrm{d}y\right], \end{aligned}$$

where $t=(x,n)$ and $t'=(x',n')$. We add and subtract $\sum_{k=1}^{\infty} w_{x',k}F(\cdot;\theta_{x,k})$ and apply the triangle inequality to find that

$$d_c(t,t') \le \mathbb{E}\left[d_{\mathrm{TV}}(F(\cdot;\theta_{x,1}),F(\cdot;\theta_{x',1}))\right] + \sum_{k=1}^{\infty}\mathbb{E}\left|w_{x,k}-w_{x',k}\right|.$$

Since $w_{x,k}$ is a product of independent random variables, Lemma 7 yields

$$\begin{aligned} d_c(t,t') \le \mathbb{E}&\left[d_{\mathrm{TV}}(F(\cdot;\theta_{x,1}),F(\cdot;\theta_{x',1}))\right] + \\ &\mathbb{E}\left[\left|v_{x,1}-v_{x',1}\right|\right]\sum_{k=1}^{\infty}\left(\left(\frac{\alpha}{\alpha+1}\right)^{k-1} + \frac{k-1}{1+\alpha}\left(\frac{\alpha}{\alpha+1}\right)^{k-2}\right). \end{aligned}$$

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

$$d_c(t,t') \le \mathbb{E}\left[d_{\mathrm{TV}}(F(\cdot;\theta_{x,1}),F(\cdot;\theta_{x',1}))\right] + C\mathbb{E}\left|v_{x,1}-v_{x',1}\right|.$$

Therefore, if the stochastic processes for the parameters and stick variables are both smooth enough such that

$$\max\left\{\mathbb{E}\left[d_{\mathrm{TV}}(F(\cdot;\theta_{x,1}),F(\cdot;\theta_{x',1}))\right], \mathbb{E}\left|v_{x,1}-v_{x',1}\right|\right\} \le (1+C)^{-1}f(d(t,t')),$$

for some function $f:\mathbb{R}_+\to\mathbb{R}_+$ such that $\lim_{x\to 0} f(x) = f(0) = 0$, then by Proposition 3 X is f-locally exchangeable. Many dependent processes (e.g., (Foti and Williamson, 2015)) similar to the dependent Dirichlet process (and kernel beta process below) can be shown to exhibit local exchangeability using similar techniques.

3.3. *Kernel beta processes.* Another example of a model exhibiting local exchangeability from the Bayesian nonparametrics literature is the kernel beta process latent feature model (Ren et al., 2011). In a typical nonparametric latent feature modelling setting, we have observations generated via

$$X_n = F\left(\cdot; Z_n\right), \qquad\qquad Z_n \overset{\text{indep}}{\sim} \mathrm{BeP}(\sum_{k=1}^{\infty} w_k \delta_{\theta_k}),$$

where $(w_k)_{k=1}^{\infty}$ are the feature frequencies satisfying $w_k \in [0,1]$, $\sum_{k=1}^{\infty} w_k < \infty$, $(\theta_k)_{k=1}^{\infty}$ are the feature parameters, BeP is the Bernoulli process that sets $Z_n(\{\theta_k\}) = 1$ with probability w_k and 0 otherwise independently across $k \in \mathbb{N}$, and F is the likelihood for each observation. A popular nonparametric prior for the weights and feature parameters is the beta process (Hjort, 1990), defined by

$$(\theta_k, w_k)_{k=1}^{\infty} \sim \mathrm{PP}(\gamma H(\mathrm{d}\theta) c(\theta) w^{-1} (1-w)^{c(\theta)-1} \mathrm{d}w),$$

where PP is a Poisson point process parametrized by its mean measure, c is some positive function, H is a probability distribution, and $\gamma > 0$. When the observations come with covariate information, the *kernel* beta process (Ren et al., 2011) may be used to capture similarities in the latent features of related populations. In particular, this model includes the modification

$$Z_{x,n} \overset{\text{indep}}{\sim} \mathrm{BeP}(\sum_{k=1}^{\infty} \kappa(x, x_k; \psi_k) w_k \delta_{\theta_k}),$$

where $\kappa(x, x_k; \psi_k)$ is a kernel function with range in $[0,1]$ centered at x_k with parameters ψ_k, and

$$(x_k, \psi_k, \theta_k, w_k)_{k=1}^{\infty} \sim \mathrm{PP}(Q(\mathrm{d}x) R(\mathrm{d}\psi) \gamma H(\mathrm{d}\theta) c(\theta) w^{-1} (1-w)^{c(\theta)-1} \mathrm{d}w),$$

where Q and R are probability distributions. In other words, the kernel beta process endows each atom with i.i.d. covariates x_k and parameters ψ_k and modifies the observation process such that observation n with covariate x is more likely to select those features k for which x_k is close to x. Taking $\mathbb{R}$ to be the space of covariates for simplicity, again we have $\mathcal{T} = \mathbb{R} \times \mathbb{N}$ and $d((x,n),(x',n')) = |x - x'|$. Marginalizing $Z_{x,n}$, the canonical pseudometric is

$$d_c(t,t') = \mathbb{E}\left[d_{\mathrm{TV}}\left(\mathbb{E}\left[F(\cdot; Z_{x,n})\right], \mathbb{E}\left[F(\cdot; Z_{x',n'})\right]\right)\right],$$

where $t = (x, n)$ and $t' = (x', n')$. Suppose F is γ-Hölder continuous in total variation for $0 < \gamma \leq 1$, $C \geq 0$ in the sense that

$$d_{\mathrm{TV}}\left(\mathbb{E}\left[F(\cdot; Z)\right], \mathbb{E}\left[F(\cdot; Z')\right]\right) \leq C\left(\sum_{k=1}^{\infty} \|\theta_k\| \left|p_k - p'_k\right|\right)^{\gamma}$$

for any collection of points $\{\theta_k\}_{k=1}^{\infty}$, where $Z(\{\theta_k\}) = 1$, $Z'(\{\theta_k\}) = 1$ independently with probability p_k and p'_k, respectively, and both assign 0 mass to all other sets. Then

$$d_c(t, t') \leq C\mathbb{E}\left[\sum_{k=1}^{\infty} |\kappa(x, x_k; \psi_k) - \kappa(x', x_k; \psi_k)| w_k \|\theta_k\|\right]^{\gamma}.$$

Finally, if the kernel κ is α-Hölder continuous with constant $C'(\psi)$ depending on ψ, the independence of θ_k, w_k, and ψ_k may be used to show that

$$\begin{aligned} d_c(t, t') &\leq C\mathbb{E}\left[\sum_{k=1}^{\infty} C'(\psi_k)|x - x'|^{\alpha} w_k \|\theta_k\|\right]^{\gamma} \\ &= C\left(\mathbb{E}\left[C'(\psi_1)\right] |x - x'|^{\alpha} \mathbb{E}\left[\|\theta_1\|\right] \mathbb{E}\left[\sum_{k=1}^{\infty} w_k\right]\right)^{\gamma}. \end{aligned}$$

Therefore by Proposition 3, the observations are f-locally exchangeable, where $f(x) = C'' x^{\alpha\gamma}$ and C'' collects the product of constants from the previous expression.

3.4. *Dynamic topic model.* The final example of local exchangeability is the dynamic topic model (Blei and Lafferty, 2006; Wang, Blei and Heckerman, 2008), a model for text data that extends latent Dirichlet allocation (Blei, Ng and Jordan, 2003) to incorporate timestamp covariate information. In a continuous version of the model, observations are generated via

$$D_{n,x} \sim \mathsf{Multi}(N, \sum_{k=1}^{K} \theta_{x,k} \pi(\beta_{x,k})), \quad \theta_x \sim \mathsf{Dir}(\pi(\alpha_x)), \quad N \sim \mathsf{Poiss}(\mu),$$

where $x \in \mathbb{R}$ represents timestamps, $\alpha_x \in \mathbb{R}^K$ is a vector of K independent Wiener processes representing the popularity of K topics at time x, $\beta_{x,k} \in \mathbb{R}^V$ is a vector of V independent Wiener processes representing the word frequencies for vocabulary of size V in topic k, π is any L-Lipschitz mapping from multidimensional real space to the probability simplex $\pi : \mathbb{R}^d \to \Delta^{d-1}$ for any $d \in \mathbb{N}$, μ is the mean number of words per document, and $D_{n,x} \in \mathbb{N}^V$

is the vector of counts of each vocabulary word in the n^{th} document observed at time x. Here the covariate space is $\mathcal{T} = \mathbb{R} \times \mathbb{N}$, and the observations are count vectors in $\mathbb{N}^V$ where V is the vocabulary size. In this setting, the canonical pseudometric is

$$d_c(t,t') = \mathbb{E}\left[d_{\text{TV}}\left(\mathsf{Multi}(N, \sum_{k=1}^{K} \theta_{x,k}\pi(\beta_{x,k})), \mathsf{Multi}(N, \sum_{k=1}^{K} \theta_{x',k}\pi(\beta_{x',k}))\right)\right],$$

where $t = (x, n)$ and $t' = (x', n')$. But since multinomial variables are a function (in particular, a sum) of a product of independent categorical random variables, Lemma 8 yields the bound

$$\leq \mathbb{E}\left[N d_{\text{TV}}\left(\mathsf{Categorical}(\sum_{k=1}^{K} \theta_{x,k}\pi(\beta_{x,k})), \mathsf{Categorical}(\sum_{k=1}^{K} \theta_{x',k}\pi(\beta_{x',k}))\right)\right].$$

We evaluate the total variation between two categorical distributions and apply the triangle inequality to find that

$$\begin{aligned}
&\leq \mathbb{E}\left[\frac{N}{2}\sum_{v=1}^{V}\left|\sum_{k=1}^{K} \theta_{x,k}\pi(\beta_{x,k})_v - \sum_{k=1}^{K} \theta_{x',k}\pi(\beta_{x',k})_v\right|\right] \\
&\leq \frac{\mu}{2}\sum_{v=1}^{V}\sum_{k=1}^{K}\mathbb{E}\left[\left|\theta_{x,k} - \theta_{x',k}\right| \pi(\beta_{x,k})_v + \theta_{x',k}\left|\pi(\beta_{x,k})_v - \pi(\beta_{x',k})_v\right|\right].
\end{aligned}$$

Since $\sum_{v=1}^{V} \pi(\beta_{x,k})_v = \sum_{k=1}^{K} \theta_{x',k} = 1$, the components of $\theta_{x,k}$ and $\beta_{x,k}$ are i.i.d. across k, and π is L-Lipschitz,

$$\begin{aligned}
&\leq \frac{\mu}{2}\left(KL\mathbb{E}\left|\alpha_{x,1} - \alpha_{x',1}\right| + VL\mathbb{E}\left|\beta_{x,1,1} - \beta_{x',1,1}\right|\right] \\
&\leq \frac{\mu L\,(K+V)}{2}\sqrt{|x-x'|},
\end{aligned}$$

where the last line follows by Jensen's inequality. Therefore the observations are f-locally exchangeable with $f(x) = \frac{1}{2}\mu L\,(K+V)\sqrt{x}$.

4. Discussion. The major question posed in this paper is whether a version of de Finetti's theorem holds when the distribution of observations is not exactly invariant under permutations. In order to answer the question, we have introduced a relaxed notion of *local* exchangeability in which swapping data associated with nearby covariates causes a bounded change in total variation distance, generalizing both partial and classical exchangeability. Many popular covariate-dependent statistical models—which violate the assumptions of classical exchangeability—have observations that are locally exchangeable. Our main results show that, indeed, de Finetti's theorem is "robust" to the real world: local exchangeability implies independence conditioned on a latent measure-valued process (Theorem 4); the smoother the local exchangeability is, the smoother the underlying process (Theorem 5); and inexact covariate information due to discretization (or "binning") has information content that degrades smoothly with the discretization coarseness (Theorem 6). This paper leaves a number of further statistical questions for future work, such as testing for local exchangeability and using local exchangeability assumptions in the design of computational procedures.

Acknowledgements. The authors thank Jonathan Huggins for illuminating discussions. 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.

APPENDIX A: ADDITIONAL TECHNICAL LEMMAS

LEMMA 7. *For any two sequences of real numbers* $(a_i)_{i=1}^\infty$, $(b_i)_{i=1}^\infty$,

$$\left|\prod_{i=1}^\infty a_i - \prod_{i=1}^\infty b_i\right| \leq \sum_{i=1}^\infty |a_i - b_i| \left(\prod_{j=1}^{i-1} b_j\right)\left(\prod_{j=i+1}^\infty a_i\right).$$

PROOF. The proof follows by adding and subtracting $b_1 \prod_{i=2}^\infty a_i$, then $b_1 b_2 \prod_{i=3}^\infty a_i$, etc., and then using the triangle inequality. □

LEMMA 8 ((Reiss, 1981)). *For any two finite product probability measures* $\mu = \mu_1 \times \cdots \times \mu_N$ *and* $\nu = \nu_1 \times \cdots \times \nu_N$,

$$1 - \exp\left(-\frac{1}{2}\sum_{n=1}^N d_{\mathrm{TV}}(\mu_n, \nu_n)^2\right) \leq d_{\mathrm{TV}}(\mu, \nu) \leq \sum_{n=1}^N d_{\mathrm{TV}}(\mu_n, \nu_n).$$

LEMMA 9. *Let* X, Y *be bounded random variables in* $[a, b]$ *for some* $a, b \in \mathbb{R}$, $a \leq b$, *and* U, V *be random elements in some probability space.*

1. *If* $\|(X,U) - (Y,U)\|_{\mathrm{TV}} \leq \epsilon$, *then* $\mathbb{E}\left|\mathbb{E}\left[X \mid U\right] - \mathbb{E}\left[Y \mid U\right]\right| \leq (b-a)\epsilon$.
2. *If* $\|(X,U) - (X,V)\|_{\mathrm{TV}} \leq \epsilon$, *then for any 1-Lipschitz function* $h : \mathbb{R} \to \mathbb{R}$, $|\mathbb{E}\left[h(\mathbb{E}\left[X \mid U\right]) - h(\mathbb{E}\left[X \mid V\right])\right]| \leq 3(b-a)\epsilon$.

Proof. 1. Denoting $Q := \mathbb{1}\left[\mathbb{E}\left[X \mid U\right] > \mathbb{E}\left[Y \mid U\right]\right]$,

$$\mathbb{E}\left|\mathbb{E}\left[X \mid U\right] - \mathbb{E}\left[Y \mid U\right]\right| = \mathbb{E}\left[\mathbb{E}\left[X \mid U\right](2Q-1) - \mathbb{E}\left[Y \mid U\right](2Q-1)\right].$$

Using the fact that Q is measurable with respect to U and the tower property yields

$$\begin{aligned}
&= \mathbb{E}\left[X\left(2Q-1\right)\right] - \mathbb{E}\left[Y\left(2Q-1\right)\right] \\
&= (b-a)\left(\mathbb{E}\left[\frac{X-a}{b-a}\left(2Q-1\right)\right] - \mathbb{E}\left[\frac{Y-a}{b-a}\left(2Q-1\right)\right]\right).
\end{aligned}$$

Since the difference is between the expectation of a function bounded in $[0,1]$ evaluated at X, U and Y, U, the assumed total variation bound provides the result.

2. First, note that $\sup_{x,y\in[a,b]}|h(x)-h(y)| \le b-a$ by 1-Lipschitz continuity. Then defining $A(U) := \mathbb{E}\left[X \mid U\right]$ and $B(V) = \mathbb{E}\left[X \mid V\right]$, the triangle inequality yields

$$\left|\mathbb{E}\left[h(A(U))\right] - \mathbb{E}\left[h(B(V))\right]\right| \le \mathbb{E}\left|A(U)-B(U)\right| + \left|\mathbb{E}\left[h(B(V)) - h(B(U))\right]\right|.$$

The right hand term is bounded by $(b-a)\epsilon$ by the assumed total variation bound and 1-Lipschitz continuity. Defining $Q(u) = \mathbb{1}\left[A(u) \ge B(u)\right]$,

$$\begin{aligned}
&\mathbb{E}\left|A(U)-B(U)\right| \\
&\quad = \mathbb{E}\left[A(U)(2Q(U)-1) - B(U)(2(Q(U)-1))\right] \\
&\quad = (b-a)\mathbb{E}\left[\frac{A(U)-a}{b-a}(2Q(U)-1) - \frac{B(V)-a}{b-a}(2Q(V)-1)\right] \\
&\quad + (b-a)\mathbb{E}\left[\frac{B(V)-a}{b-a}(2(Q(V)-1) - \frac{B(U)-a}{b-a}(2(Q(U)-1))\right].
\end{aligned}$$

The first term in the expression can be bounded by $(b-a)\epsilon$ via substitution of the conditional expectation formulae for A, B, using the tower property, and controlling the difference in expectations with the assumed total variation bound. The second term is again a difference in expectation of a bounded function under U and V with the same bound $(b-a)\epsilon$. □

REFERENCES

Aldous, D. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* **11** 581–598.

Aldous, D. (1985). *Exchangeability and related topics. École d'été de probabilités de Saint-Flour , XIII.* Springer, Berlin.

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

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

Borgs, C., Chayes, J., Cohn, H. and Holden, N. (2018). Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research* **18** 1–71.

Broderick, T., Pitman, J. and Jordan, M. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis* **8** 801–836.

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. *The Annals of Statistics* **47** 67–92.

Campbell, T., Cai, D. and Broderick, T. (2018). Exchangeable trait allocations. *Electronic Journal of Statistics (in press).*

Caron, F. and Fox, E. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society Series B* **79** 1295–1366.

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

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

Crane, H. and Dempsey, W. (2016b). Relational Exchangeability. *arXiv:1607.06762v1.*

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. *Annales de l'Institut Henri Poincaré* **7** 1–68.

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.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese* **36** 271–281.

Diaconis, P. (1988). Recent progress on de Finetti's notions of exchangeability. *Bayesian Statistics* **3** 111–125.

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

Diaconis, P. and Freedman, D. (1980a). Finite exchangeable sequences. *The Annals of Probability* **8** 745–764.

Diaconis, P. and Freedman, D. (1980b). De Finetti's theorem for Markov chains. *The Annals of Probability* **8** 115–130.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230.

Foti, N. and Williamson, S. (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 359–371.

Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70** 419–435.

Halmos, P. (1978). 2^{nd} ed. *Graduate Texts in Mathematics*. Springer, New York.

Hewitt, E. and Savage, L. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society* **80** 470–501.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* **18** 1259–1294.

Hoover, D. (1979). Relations on probability spaces and arrays of random variables

Technical Report, Institute for Advanced Study, Princeton University.

Janson, S. (2017). On edge exchangeable random graphs. *Journal of Statistical Physics*.

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.

Kallenberg, O. (1990). Exchangeable random measures in the plane. *Journal of Theoretical Probability* **3** 81–136.

Kallenberg, O. (2002). *Foundations of modern probability*, 2^{nd} ed. *Probability and its applications*. Springer, New York.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*, 1^{st} ed. *Probability and its applications*. Springer-Verlag, New York.

Kingman, J. F. C. (1978). The representation of partition structures. *Journal of the London Mathematical Society* **2** 374–380.

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

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

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

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

Orbanz, P. and Roy, D. (2015). Bayesian models of graphs, arrays, and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 437–461.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102** 145–158.

Pothoff, J. (2009). Sample properties of random fields II: continuity. *Communications on Stochastic Analysis* **3** 331–348.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Reiss, R.-D. (1981). Approximation of product measures with an application to order statistics. *The Annals of Probability* **9** 335–341.

Ren, L., Wang, Y., Dunson, D. and Carin, L. (2011). The kernel beta process. In *Advances in Neural Information Processing Systems*.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–350.

Veitch, V. and Roy, D. (2015). The class of random graphs arising from exchangeable random measures. *arXiv:1512.03099*.

Wang, C., Blei, D. and Heckerman, D. (2008). Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence*.

Department of Statistics
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
E-mail: trevor@stat.ubc.ca, saif.syed@stat.ubc.ca

Department Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA, USA 94720
E-mail: chiaoyu@berkeley.edu, jordan@eecs.berkeley.edu

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, USA 02139
E-mail: tbroderick@csail.mit.edu