Electronic Journal of Statistics
Vol. 3 (2009) 290–315
ISSN: 1935-7524
DOI: 10.1214/08-EJS269

Building hyper Dirichlet processes for graphical models

Daniel Heinz

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
e-mail: dheinz@stat.cmu.edu

Abstract: Graphical models are used to describe the conditional independence relations in multivariate data. They have been used for a variety of problems, including log-linear models (Liu and Massam, 2006), network analysis (Holland and Leinhardt, 1981; Strauss and Ikeda, 1990; Wasserman and Pattison, 1996; Pattison and Wasserman, 1999; Robins et al., 1999), graphical Gaussian models (Roverato and Whittaker, 1998; Giudici and Green, 1999; Marrelec and Benali, 2006), and genetics (Dobra et al., 2004). A distribution that satisfies the conditional independence structure of a graph is *Markov*. A graphical model is a family of distributions that is restricted to be Markov with respect to a certain graph. In a Bayesian problem, one may specify a prior over the graphical model. Such a prior is called a *hyper Markov law* if the random marginals also satisfy the independence constraints. Previous work in this area includes (Dempster, 1972; Dawid and Lauritzen, 1993; Giudici and Green, 1999; Letac and Massam, 2007). We explore graphical models based on a non-parametric family of distributions, developed from Dirichlet processes.

AMS 2000 subject classifications: Primary 36E05; secondary 62G99.
Keywords and phrases: Hyper Markov law, stick-breaking measure, non-parametric prior, decomposable graphical distribution, covariance selection.

Received July 2008.

Contents

1. Introduction

Markov distributions are multivariate measures that satisfy a specified set of conditional independence relations. Often, an undirected graph is useful to represent this struture. A measure is Markov with respect to a graph if whenever two variables have no edge between them, they are conditionally independent given the other variables in the graph. Graphical models are especially important in many high-dimensional problems to alleviate computational burden or to solve under-identified models. An example of the former issue arises in Gaussian models in which we may want to invert the $p \times p$ covariance matrix. Graphical models help by reducing the complexity of this algorithm. Under-identified models arise in modern statistical fields such as machine learning (REF) and biostatistics (REF) under the moniker of "$n \ll Ap$ problems". Graphical models are one way to reduce the effective dimensionality of the problem and identify a solutions. Dawid and Lauritzen (1993) extended the notion of Markovity to the parameter space. In Bayesian statistics, the measure of the data is random, and therefore has its own distribution called the prior. A prior law over Markov measures is *hyper Markov* if it gives probability one to Markov measures and the random marginal measures have the specified conditional independence structure. An example is the hyper inverse Wishart distribution, which serves as a prior for the multivariate Gaussian with known mean. The usual inverse Wishart is a specific case, which is hyper Markov for the saturated model.

Like all parametric models, the hyper inverse Wishart prior makes strong assumptions about the shape of the distribution. In many applications, such assumptions are undesirable. Thus, there is a need for a model which takes advantage of graphs wihtout relying on parametric models. The current paper aims to answer this need by applying a non-parametric approach to graphical models. In contrast to parametric models, non-parametric models make weak assumptions. Typical assumptions include continuity and the existence of some number of derivatives. For example, one may specify that the distribution is smooth, having derivatives of all orders. We extend the framework laid by Dawid and Lauritzen to develop the *hyper Dirichlet process*. We begin with the Dirichlet Process, a commonly used non-parametric prior law. We then describe how to build this family into a non-parametric hyper Markov prior.

As in Dawid and Lauritzen (1993), we restrict our attention to decomposable graphs. The benefit of this is that a decomposable graph can be easily built up from smaller components called cliques which intersect to form the entire graph.

Dawid and Lauritzen begin by considering a base distribution for each clique. The only requirement is that these distributions agree where the cliques intersect. They weave together base distributions by taking the base measure of one clique as its marginal, and conditioning the second clique on the intersection. They repeat this process for every clique. The third clique is added by conditioning its base measure on the intersection with the previous two cliques. This process is repeated until all the cliques have been combined. The end result is a Markov distribution whose marginal over each clique is the clique's base. For a prior on Markov distributions, Dawid and Lauritzen construct a hyper Markov law in the same way.

As an example of the Dawid and Lauritzen (1993) construction, consider the problem of estimating the covariance matrix of a multivariate Gaussian. If we believe that the data exhibit some conditional independence structure, this implies certain constraints on the covariance matrix. (Speed and Kiiveri, 1986) showed that the sufficient statistics are the component covariance matrices belonging to each clique. The inverse Wishart is the usual prior for the saturated model which has no constraints on the covariance matrix. In a non-saturated model, the sub-matrix of each clique is unconstrained, except that the sub-matrices must agree where their indices intersect. For this reason, the inverse Wishart is the natural choice as the base measure for each clique. The sub-matrix for the first clique has an inverse Wishart prior. If the graph is connected and the cliques have a perfect ordering (see Section 2.2), then the first and second sub-matrices have some elements in common. Thus, the sub-matrix for the second clique is the inverse Wishart, conditional on knowing some of the elements. By repeating the conditioning for each clique, we define the *hyper inverse Wishart*.

In the current paper, we apply this framework to non-parametric priors. Instead of the inverse Wishart, the Dirichlet process prior is the base measure for each clique. Following the analogy, we build the marginals into a hyper Markov prior, which we refer to as the hyper Dirichlet process. The Dirichlet process is a special case of tail-free processes (Ferguson, 1973). Dirichlet processes have been used for non-parametric priors in many areas, including block models (Bush and MacEachern, 1996), survival analysis (Susarla and Ryzin, 1976; Ghosh and Ramamoorthi, 1995; Kim and Lee, 2001), and non-stationary point processes (Pievatolo and Rotondi, 2000). These are all areas that could potentially use a hyper Dirichlet process in multidimensional problems. In Section 2, we explain notation and formalize some of the ideas presented so far. In Section 3, we describe the Dirichlet and some previous results. In Section 4 we weave Dirichlet processes on the cliques to build the hyper Dirichlet process, and show that it is a hyper Markov prior. Finally, we explore applications for this framework in Section 5.

2. Notation and setting

2.1. Basic graph theory

Throughout this paper we consider a graph, $\mathcal{G}$, with vertex set $\mathbf{V}$ and edge set $\mathbf{E}$. By convention, we assume that $(\gamma, \gamma) \in \mathbf{E}$ for all γ. We call such edges *loops*.

There is no practical difference if loops are excluded from $\mathbf{E}$, though some minor changes are required for certain definitions. If $\mathbf{A} \subseteq \mathbf{V}$, then $\mathcal{G}_\mathbf{A}$ is the subgraph of $\mathcal{G}$ over $\mathbf{A}$. The subgraph $\mathcal{G}_\mathbf{A}$ has vertex set $\mathbf{A}$, and edge set $\mathbf{E}_\mathbf{A} = (\mathbf{A} \times \mathbf{A}) \cap \mathbf{E}$. We say that $\mathbf{A}$ induces the subgraph $\mathcal{G}_\mathbf{A}$. If $\mathbf{E}_\mathbf{A} = \mathbf{A} \times \mathbf{A}$, then $\mathcal{G}_\mathbf{A}$ is complete. A *clique* is a set $\mathbf{A}$ such that $\mathcal{G}_\mathbf{A}$ is complete and for any superset $\mathbf{B} \supset \mathbf{A}$, $\mathcal{G}_\mathbf{B}$ is not complete. For example, if $\mathcal{G}$ itself is complete, then there is one clique, viz. $\mathbf{V}$.

A k-path is a sequence $(\gamma_0, \gamma_1, \ldots, \gamma_k)$, such that $(\gamma_i, \gamma_{i+1}) \in \mathbf{E}$, for $0 \leq i < k$. If $\mathbf{A}$ and $\mathbf{B}$ are subsets of $\mathbf{V}$, then a path between them is a path between any $a \in \mathbf{A}$ and any $b \in \mathbf{B}$. A graph is *connected* if there exists a path between every pair of subsets. A third subset $\mathbf{C} \in \mathbf{V}$ is said to *separate* $\mathbf{A}$ and $\mathbf{B}$ if every path between them contains an element of $\mathbf{C}$. A k-cycle is a path such that $k \geq 3$, $\gamma_0 = \gamma_k$ and the other elements are distinct. Within a k-cycle, a *chord* is an edge (or "short cut") between two non-consecutive nodes. A graph is *decomposable* if every cycle longer than length 3 contains a chord. A decomposable graph admits a *perfect ordering* of its cliques.

Definition 1 (Perfect Ordering**).** Suppose a graph $\mathcal{G}$ has n cliques. Let the cliques have an arbitrary ordering $\mathbf{C}_1, \ldots, \mathbf{C}_n$. Define $\mathbf{H}_k = \cup_{i=1}^k \mathbf{C}_i$. For $k \geq 2$ define $\mathbf{S}_k = \mathbf{C}_k \cap H_{k-1}$ and $\mathbf{R}_k = \mathbf{C}_k \setminus H_{k-1}$. The ordering of the cliques is a *perfect ordering* if for each $2 \leq k \leq n$, there exists $j_k < k$ such that $\mathbf{S}_k \subset \mathbf{C}_{j_k}$.

The sets $\mathbf{H}_k$ are called the histories. The separators, $\mathbf{S}_k$, separate $\mathbf{C}_k$ from the previous history. The sets $\mathbf{R}_k$ are called the residuals, which represent the new nodes being added to the history. In a perfect ordering, each new clique is separated from the current set of nodes by a single one of the earlier cliques.

For every, $\gamma \in \mathbf{V}$, X_γ is a random variable taking values in the space $(\mathcal{X}_\gamma, \mathcal{F}_\gamma)$. In this sense, we consider $\mathbf{V}$ an index set of components of some random variable $X = (X_\gamma : \gamma \in \mathbf{V})$. We denote the range and σ-field of X by $(\mathcal{X}, \mathcal{F}) = (\times_{\gamma \in \mathbf{V}} \mathcal{X}_\gamma, \times_{\gamma \in \mathbf{V}} \mathcal{F}_\gamma)$. Furthermore, we extend these definitions to subsets, $\mathbf{A} \subseteq \mathbf{V}$.

$$\begin{aligned} X_\mathbf{A} &= (X_\mathbf{A} : \gamma \in \mathbf{A}) \\ \mathcal{X}_\mathbf{A} &= \times_{\gamma \in \mathbf{A}} \mathcal{X}_\gamma \\ \mathcal{F}_\mathbf{A} &= \times_{\gamma \in \mathbf{A}} \mathcal{F}_\gamma \end{aligned}$$

Let α be a measure over some $\mathcal{X}_\mathbf{A}$, then $\overline{\alpha} = \alpha / \alpha(\mathcal{X}_\mathbf{A})$. In other words, $\overline{\alpha}$ is the probability measure proportional to α. If $\mathbf{B} \subseteq \mathbf{A}$, then $\alpha_\mathbf{B}$ is the marginal of α over $\mathcal{X}_\mathbf{B}$. Thus, $\alpha_\mathbf{B}(U) = \alpha(U \times \mathcal{X}_{\mathbf{A} \setminus \mathbf{B}})$, $\forall\ U \in \mathcal{F}_\mathbf{B}$. If α and β are both measures on some space $(\mathcal{X}, \mathcal{F})$, then we define their sum, $\alpha + \beta$, by

$$[\alpha + \beta](U) = \alpha(U) + \beta(U), \quad \forall\ U \in \mathcal{F}.$$

If $x \in \mathcal{X}$, then the delta measure δ_x is a point mass concentrated at x:

$$\delta_x(U) = \begin{cases} 1, & x \in U \\ 0, & x \notin U \end{cases}, \quad \forall U \in \mathcal{F}.$$

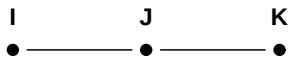

FIG 1. *A graph depicting conditional independence of I and K given J.*

2.2. Graph selection

For the remainder of the paper, we consider undirected graphs, which implies that $(i, j) \in \mathbf{E}$ if and only if $(j, i) \in \mathbf{E}$. We also assume that the graph is connected and decomposable. An undirected graph depicts the conditional independence structure for some variable X in the following sense:

$$X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} \mid X_{\mathbf{C}} \text{ whenever } \mathbf{C} \text{ separates } \mathbf{A} \text{ and } \mathbf{B}. \tag{1}$$

Definition 2 (MARKOV PROBABILITY MEASURE**).** If θ is a probability measure on $(\mathcal{X}, \mathcal{F})$, we say it is Markov on a decomposable graph, $\mathcal{G}$, if $X \sim \theta$ satisfies the conditional independences in $\mathcal{G}$.

Example 1. Let $\mathcal{G}$ be the graph depicted in Figure 1. A measure θ is Markov on $\mathcal{G}$, if and only if $X_I \perp\!\!\!\perp X_K | X_J$ whenever $X \sim \theta$.

Implicit in the definition is the fact that it is only sensible to refer to a measure as Markov in relation to a specific graph. For example, if the measure θ is not Markov on $\mathcal{G}$ in Example 1, it is still Markov on the saturated graph with $\mathbf{V} = \{I, J, K\}$. All measures over $\mathcal{X}_{\mathbf{V}}$ are trivially Markov on the saturated graph since there are no constraints on conditional independence. Furthermore, if μ is a measure such that each X_γ is independent, then it is Markov on any graph (with the appropriate vertex set.) We denote the set of all distributions that are Markov on $\mathcal{G}$ by $\mathscr{M}(\mathcal{G})$.

It will be useful to keep Figure 1 in mind throughout this paper. While the graph technically has only three variables, it is representative of any connected graph of two cliques. Instead of one variable, imagine I, J, and K to contain multiple variables, with J being the variables that belong to both cliques. I is the set of variables in one clique but not the other, and K vice versa.

Let $X \sim P \in \mathscr{F}$ be a random variable whose distribution is modeled by some family of probability distributions. In some applications, the focus is not on determining P, but on discovering the independence structure of X. A graph of this structure, $\mathcal{G}$, denotes the belief that P is Markov with respect to $\mathcal{G}$. Thus, it restricts the model to a sub-family, $\mathscr{F}_{\mathcal{G}} = \mathscr{F} \cap \mathscr{M}(\mathcal{G})$. Graph selection is the problem of determining the smallest $\mathscr{F}_{\mathcal{G}}$ that contains P. The most prevalent examples are graphical Gaussian models. Graph selection for Gaussian models is often called covariance selection. In this setting, the relevant family is the set of p-variate Gaussian distributions. Denote this family $\mathscr{N} = \{N_p(\mu, \Sigma) : \mu \in \mathcal{R}^p, \Sigma \in M_p^+\}$, where M_p^+ is the cone of real-valued, symmetric $p \times p$ matrices that are positive definite. Specifying a graph, $\mathcal{G}$, translates to putting constraints on Σ. For example, if (x_1, x_2, x_3) is such that $x_1 \perp\!\!\!\perp x_3 | x_2$, then σ_{13} is no longer a free parameter, but a function of $\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}$, and σ_{23}.

Denote the sub-family of Gaussian distributions Markov on $\mathcal{G}$ by $\mathcal{N}_{\mathcal{G}}$. Let $P_{\mathcal{G}}$ be the set of positive definite matrices such that $K_{ij} = 0$ for all $(i, j) \notin \mathbf{E}$; let $Q_{\mathcal{G}}$ be the image of $P_{\mathcal{G}}$ under matrix inversion. Speed and Kiiveri (1986) showed that if $N_p(\mu, \Sigma)$ is Markov with respect to $\mathcal{G}$, then $\Sigma \in Q_{\mathcal{G}}$. Thus $\mathcal{N}_{\mathcal{G}} = \{N_p(\mu, \Sigma) : \mu \in \mathcal{R}^p, \Sigma \in Q_{\mathcal{G}}\}$. The goal of covariance selection is to find the smallest $Q_{\mathcal{G}}$ containing Σ, the population covariance matrix.

Much progress has been made with graph selection for parametric models. Dawid and Lauritzen (1993) proved many results for decomposable graphical models, including multinomial and multivariate Gaussian problems. For example, they present the distribution of the restricted maximum likelihood estimate of Σ for the $\mathcal{N}_{\mathcal{G}}$ model with μ known. This distribution is called the hyper Wishart distribution; if we restrict $\mathcal{G}$ to the complete graph, then we recover the Wishart distribution for the measure of the MLE. Letac and Massam (2007) have extended the hyper Wishart to a richer family of distributions on $Q_{\mathcal{G}}$ and $P_{\mathcal{G}}$. Giudici and Green (1999) implemented a reversible jump Markov chain Monte Carlo algorithm for determining $\mathcal{G}$.

The family of hyper inverse Wishart distributions is the subset of Markov distributions such that each clique marginal is inverse Wishart. Carvalho et al. (2007) provide an algorithm for generating random variables from this family. For decomposable models, the presence of a perfect ordering simplifies the process. For two cliques, the algorithm begins by generating an inverse Wishart variable on one clique. If the cliques overlap, this determines some of the parameters for the other clique. Therefore, one needs to generate a conditional Wishart variable given those entries. For multiple cliques, one simply repeats this process. With a perfect ordering, the process is simplified because each new clique is conditioned on only one previous clique. Conditioning on multiple cliques can lead to moderate complications in the conditional distribution. Hence, decomposable models are computationally convenient.

3. The Dirichlet process

The Dirichlet process (Ferguson, 1973) is a special case of tail-free distributions. It is a prior, meaning that it provides a distribution over the space of probability distributions on $(\mathcal{X}, \mathcal{F})$. In this paper, we use the term *law* to refer to distributions over probability measures. However, this terminology is merely a convenience; the words "law" and "distribution" are typically interchangeable. The Dirichlet process is an example of a non-parametric law, which means that it cannot be specified by a finite-dimensional parameter. In this section, the Dirichlet process is introduced and some of its useful properties are given. This leads into the next section, in which we show how to compose a hyper Dirichlet processes from multiple Dirichlet processes.

3.1. The Ferguson (1973) Dirichlet process

Definition 3 (Dirichlet Process)**.** Let $\mathbf{A}$ be any subset of $\mathbf{V}$. Let α be a measure over $(\mathcal{X}_{\mathbf{A}}, \mathcal{F}_{\mathbf{A}})$, and let θ be a random probability measure over the

same space. We say that the distribution of θ is a *Dirichlet process* with base measure α, and write $\theta \sim DP_\alpha$, if

$$(\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots, \mathbb{P}(A_k)) \sim \mathrm{Dir}(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k)), \tag{2}$$

whenever $(A_i)_{i=1}^k$ is a partition of $\mathbf{A}$.

This definition leads to some useful properties, described in the following theorem.

Theorem 4 (Posterior Dirichlet Process). *Let $\theta \sim DP_\alpha$ and, given θ, let $X_1, \ldots, X_n$ be an iid sample from θ.*

(i) $X_i \sim \overline{\alpha} \ \ \forall\ i$.
(ii) $\theta|(X_1, \ldots, X_n) \sim DP_{\alpha'}$, *where* $\alpha' = \alpha + \sum_{i=1}^n \delta_{X_i}$.

See Theorem 1.9.4 of Schervish (1995), p. 54.

The first property states that if the random measure is integrated out, the marginal distribution of the data is $\overline{\alpha}$. This property ensures that a Markov base measure implies that the Dirichlet process, integrated over all possible θ, is a Markov distribution. This does not guarantee that the process is a hyper Markov law. That requires the stronger condition that $\theta \sim DP_\alpha$ is a Markov distribution with probability one. The second property states that if a Dirichlet process is used as a prior measure, then the posterior measure is also a Dirichlet process, with an easily updated base measure. This fact helps determine which properties of a prior will persist in the posterior.

If the prior law of θ is a Dirichlet process, then the various marginal distributions of θ will also have a Dirichlet process law. This is expressed in the following theorem.

Theorem 5 (Marginal of a Dirichlet Process). *Let $\theta \sim DP_\alpha$ be a random probability measure on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$. For $\mathbf{B} \subseteq \mathbf{A}$, the marginal of θ over $\mathbf{B}$ is $\theta_\mathbf{B} \sim DP_{\alpha_\mathbf{B}}$.*

Proof. Define $\mathbf{A}' = \mathbf{A} \setminus \mathbf{B}$. Let $B_1, B_2, \ldots, B_k$ be a measurable partition of $\mathbf{B}$.

$$\begin{aligned}(\mathbb{P}_{\theta_\mathbf{B}}(B_1), \ldots, \mathbb{P}_{\theta_\mathbf{B}}(B_k)) &= (\mathbb{P}_\theta(B_1 \times \mathcal{X}_{\mathbf{A}'}), \ldots, \mathbb{P}_\theta(B_k \times \mathcal{X}_{\mathbf{A}'})) \\ &\sim \mathrm{Dir}(\alpha(B_1 \times \mathcal{X}_{\mathbf{A}'}), \ldots, \alpha(B_k \times \mathcal{X}_{\mathbf{A}'})) \\ &= \mathrm{Dir}(\alpha_\mathbf{B}(B_1), \ldots, \alpha_\mathbf{B}(B_k)).\end{aligned}$$

□

3.2. The Dirichlet process as a prior

We proceed by showing how the Dirichlet process can be used as a non-parametric prior (see Ferguson (1973) for details.) Let F be an unknown cumulative probability distribution that we wish to estimate. For simplicity, we consider a one-dimensional random variable. Let $\pi = DP_\alpha$ be the prior law. Let the loss function be a squared error loss. Then the Bayes' risk is

$$R_\pi(F, \widetilde{F}) = \int E\big(F(t) - \widetilde{F}(t)\big)^2 dt. \tag{3}$$

The risk is minimized by setting $\widetilde{F}(t)$ to $EF(t)$, where the expectation is relative to the posterior distribution. If we observe data $X_1, X_2, \ldots X_n$, then the posterior is $DP_{\alpha'}$, where $\alpha' = \alpha + \sum_i \delta_{X_i}$ (see Theorem 7.) The posterior distribution of $P((-\infty, t])$ is Beta$(\alpha'((-\infty, t]), \alpha'((t, \infty))$. Therefore,

$$EF(t) = \frac{\alpha'((-\infty, t])}{\alpha'((-\infty, t]) + \alpha'((t, \infty))} \tag{4}$$

$$= \frac{\alpha((-\infty, t]) + \sum_{i=1}^{n} 1_{(X_i \leq t)}}{\alpha(\mathcal{X}) + n}. \tag{5}$$

The Bayes estimate can be written as a weighted sum of two estimates

$$\widetilde{F}(t) = EF(t) \tag{6}$$

$$= (1 - w)\overline{\alpha}((-\infty, t]) + w\widehat{F}(t), \tag{7}$$

where $\overline{\alpha}((-\infty, t])$ is the prior estimate, $\widehat{F}(t)$ is the empirical cdf, and $w = n(\alpha(\mathcal{X}) + n)^{-1}$ is the weight of the data. This convex combination of a prior estimate and frequentist estimate is common in Bayesian analysis. This shows the role of the base measure in the Dirichlet process. $\overline{\alpha}$ is the prior guess about the shape of the unknown distribution. $\alpha(\mathcal{X})$ is mathematically equivalent to the prior sample size.

3.3. The Dirichlet process as a stick-breaking prior

A stick-breaking process is an almost surely discrete random probability measure, θ, that can be expressed as

$$\theta(\cdot) = \sum_{k=1}^{N} w_k \delta_{Z_k}(\cdot), \tag{8}$$

where the Z_k are independently distributed atoms from some distribution H, and $\sum_{k=1}^{N} w_k = 1$ almost surely. The number of atoms, N, may be finite or infinite. The weights are determined by successively breaking random pieces of a unit-length stick. Thus, $w_1 = p_1$, $w_2 = (1 - p_1)p_2$, and $w_k = p_k \prod_{i=1}^{k-1}(1 - p_i)$. For finite N, w_N is defined by $1 - \sum_{i=1}^{N-1} w_i$, or equivalently by $\prod_{i=1}(1 - p_i)$. Traditionally, stick-breaking measures are defined such that p_k is defined as a Beta(a_k, b_k) random variable for $1 \leq k < N$. Thus, a stick-breaking measure is specified by a probability distribution P, and a countable sequence of Beta parameters $(a_k, b_k)_{k=1}^{N-1}$. Sethuraman (1994) showed that a Dirichlet Process is a stick-breaking measure with $Z_k \sim \overline{\alpha}$, and $(a_k, b_k) = (0, \alpha(\mathcal{X}))$ for all $k \in \mathbb{N}$. This relationship leads to an alternative definition of the Dirichlet process.

Definition 6 (Dirichlet Process (alternate definition)**).** Let $\mathbf{A}$ be any subset of $\mathbf{V}$. Let H be a probability measure on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$, and let θ be a random probability measure over the same space. For $\nu > 0$, we say that the distribution

of θ is a *Dirichlet process* with base distribution (or measure) H and precision ν, and write $\theta \sim DP(\nu H)$, if

$$(\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots, \mathbb{P}(A_k)) \sim \text{Dir}(\nu H(A_1), \nu H(A_2), \ldots, \nu H(A_k)), \tag{9}$$

whenever $(A_i)_{i=1}^k$ is a partition of $\mathbf{A}$.

Note that this distribution is equivalent to Definition 3 by letting $\alpha = \nu H$. For example, ν is equivalent to the prior sample size, and H is equivalent to the prior mean. In this definition, ν and H are easily translated as the parameters of a stick-breaking measure. That is, the random atoms are iid H, and $p_k \sim \text{Beta}(0, \nu)$ for all $k \in \mathbb{N}$. Because the stick-breaking representation is useful for many of the theorems we prove, Definition 6 will be the definition of choice for much of the current paper.

The previous theorems regarding Dirichlet processes can be expressed using νH notation. For example, we rewrite Theorem 4 regarding the posterior Dirichlet process.

Theorem 7 (Posterior Dirichlet Process (alternate)**).** *Let $\theta \sim DP(\nu H)$ and, given θ, let $X_1, \ldots, X_n$ be an iid sample from θ.*

(i) $X_i \sim H \quad \forall\ i$.
(ii) $\theta|(X_1, \ldots, X_n) \sim DP(\nu' H')$, *where* $\nu' = \nu + n$ *and* $H' = (\nu + n)^{-1}(\nu P + \sum_{i=1}^n \delta_{X_i})$.

In the following section we introduce the hyper Dirichlet process and show that it is an example of a stick-breaking measure. We then use Equation 8 to prove some of its properties. While we focus on the hyper Dirichlet Process for simplicity and concreteness, many of the results apply to other stick-breaking processes as well.

4. The hyper Dirichlet process

Consider a multivariate variable X with distribution θ. Suppose that we know little about θ, other than it is Markov on some decomposable graph, $\mathcal{G}$. In this case we may wish to specify a non-parametric prior for θ. For example, we focus on the Dirichlet process. There are two main difficulties with this approach. The first is the elicitation of a proper base measure. The second is ensuring that the Dirichlet process gives probability one to $\mathscr{M}(\mathcal{G})$. Both concerns are addressed by using a framework that we dub the *hyper Dirichlet process*.

To define a hyper Dirichlet process, we begin by eliciting a base measure for each clique in $\mathcal{G}$. Hopefully, this is simpler than eliciting a base measure for the entire graph at once. These base measures are combined to form a base measure over the entire graph. We define these combinations in a way which will ensure that the support of the process lies within the set of Markov distributions on $\mathcal{G}$. In the remainder of this section, we provide details to this method and show that it satisfies the Markov property.

4.1. Markov combinations of probability measures

Dawid and Lauritzen (1993) show that if two subsets of $\mathbf{V}$ are each endowed with a marginal probability measure, then there is a logical choice for their joint distribution, provided the marginals satisfy a consistency condition.

Definition 8 (Consistency (of probability measures)**).** Suppose $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$. Let μ and λ be probability measures on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$ and $(\mathcal{X}_\mathbf{B}, \mathcal{F}_\mathbf{B})$, respectively. We say that μ and λ are *consistent* if they induce the same marginal over $\mathcal{X}_{\mathbf{A}\cap\mathbf{B}}$.

Note that μ and λ are consistent only if

$$\mu(\mathcal{X}_{\mathbf{A}\setminus\mathbf{B}} \times U) = \lambda(\mathcal{X}_{\mathbf{B}\setminus\mathbf{A}} \times U) \ \ \forall\, U \in \mathcal{F}_{\mathbf{A}\cap\mathbf{B}}.$$

Theorem 9. *Suppose μ on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$ and λ on $(\mathcal{X}_\mathbf{B}, \mathcal{F}_\mathbf{B})$ are consistent probability measures, with $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$. There exists an almost-everywhere unique distribution, α, such that:*

(i) $\alpha_\mathbf{A} = \mu$,
(ii) $\alpha_\mathbf{B} = \lambda$.
(iii) $\alpha \in \mathscr{M}(\mathcal{G}_{\mathbf{A}\cup\mathbf{B}})$,

Proof. Construct α such that its marginal over $\mathcal{X}_\mathbf{A}$ is μ, so that condition (i) is satisfied. Specify its conditional distributions over $\mathcal{X}_\mathbf{B}$ given $X_\mathbf{A}$ to be the same as the conditional distributions of λ given $X_{\mathbf{A}\cap\mathbf{B}}$. This ensures that (iii) holds as well. Let $\mathbf{C} = \mathbf{A} \cap \mathbf{B}$ and $\mathbf{B}' = \mathbf{B} \setminus \mathbf{A}$. Then for any $U \in \mathbf{B}'$ and $V \in \mathbf{C}$,

$$\begin{aligned}
\mathbb{P}_{\alpha_\mathbf{B}}(U \times V) &= \mathbb{P}_{\alpha_{\mathbf{B}'|\mathbf{C}}}(U|V)\mathbb{P}_{\alpha_\mathbf{C}}(V) \\
&= \mathbb{P}_{\lambda_{\mathbf{B}'|\mathbf{C}}}(U|V)\mathbb{P}_{\mu_\mathbf{C}}(V) \\
&= \mathbb{P}_{\lambda_{\mathbf{B}'|\mathbf{C}}}(U|V)\mathbb{P}_{\lambda_\mathbf{C}}(V) \\
&= \mathbb{P}_\lambda(U \times V).
\end{aligned}$$

The second equation follows from the construction of α. The third equation is ensured since μ and λ are consistent. Hence, condition (ii) is also satisfied. Furthermore, the conditional distributions are unique, except over some subset of $\mathcal{X}_\mathbf{C}$ with zero measure under λ, and hence also under μ by consistency. Therefore, this construction gives (a version of) the unique distribution satisfying the conditions. □

Definition 10 (Markov Combination (of probability measures)**).** Let μ and λ be as in Theorem 9. We call the unique distribution satisfying (i)-(iii) the *Markov Combination* of μ and λ, and denote it by $\mu \star \lambda$.

Now suppose $\mathcal{G}$ has a perfect ordering of cliques $(\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k)$, and that each clique $\mathbf{C}_i$ is imbued with a marginal probability distribution Q_i. Further suppose that Q_i and Q_j are consistent for every pair (i, j). Each clique is consistent with the previous history regarding the separator, since the separator is contained by a single previous clique. Using the idea of a Markov combination iteratively, we stitch together a distribution that is Markov on $\mathcal{G}$ and has the given

marginals. Define $H_1 = Q_1$, and $H_i = H_{i-1} \star Q_i$ for $i \geq 2$. Dawid and Lauritzen show that $H = H_k$ is the unique Markov distribution satisfying $H_{\mathbf{C}_k} = Q_k$. We call H the Markov combination of $Q_1, \ldots, Q_k$. In general, we may write $\star(Q_1, \ldots, Q_k)$ to indicate a Markov combination with the understanding that the cliques are perfectly ordered and $Q_1, \ldots, Q_k$ are pairwise consistent.

4.2. Markov combinations of finite measures

Using Markov combinations, we are able to take probability distributions and build a distribution over the entire graph. The base measure of a Dirichlet process, however, is not necessarily a probability distribution. Therefore, we proceed by extending Markov combinations to finite measures in general. For probability measures, we required the conditionals over $\mathcal{X}_{\mathbf{A}\cap\mathbf{B}}$ to be the same. We simply extend this definition to any finite measure.

Definition 11 (Consistency of Finite Measures**).** Let μ be a finite measure over $(\mathcal{X}_{\mathbf{A}}, \mathcal{F}_{\mathbf{A}})$ and λ be a finite measure over $(\mathcal{X}_{\mathbf{B}}, \mathcal{F}_{\mathbf{A}})$. We say that μ and λ are consistent if they induce the same marginal measure over $\mathbf{A} \cap \mathbf{B}$. That is, μ and λ are consistent if

$$\mu(\mathcal{X}_{\mathbf{A}\setminus\mathbf{B}} \times U) = \lambda(\mathcal{X}_{\mathbf{B}\setminus\mathbf{A}} \times U) \quad \forall\, U \in \mathcal{F}_{\mathbf{A}\cap\mathbf{B}}. \tag{10}$$

Recall that $\overline{\mu}$ is the probability measure proportional to μ. Equation 10 holds if the following two conditions are satisfied:

1. $\overline{\mu}$ and $\overline{\lambda}$ are consistent.
2. $\mu(X_{\mathbf{A}}) = \lambda(X_{\mathbf{B}})$.

Consider these two conditions in the context of base measures for Dirichlet processes. $\overline{\mu}$ is the prior guess about the probability distribution of $X_{\mathbf{A}}$, and $\overline{\lambda}$ is the prior guess for $X_{\mathbf{B}}$. The first condition therefore states that the priors must agree about the distribution of $X_{\mathbf{A}\cap\mathbf{B}}$. It is reasonable to require that our prior is coherent in this way. The second condition states that the prior sample sizes for both sets of variables must be equal. This restraint is perhaps less desirable. It would be perfectly logical to be more certain about certain dimensions than others. Unfortunately, any measure on $\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}$ must satisfy

$$\alpha(\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}) = \int_{\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}} d\alpha = \int_{\mathcal{X}_{\mathbf{A}}} \int_{\mathcal{X}_{\mathbf{B}\setminus\mathbf{A}}} d\alpha = \int_{\mathcal{X}_{\mathbf{A}}} d\alpha_{\mathbf{A}} = \alpha_{\mathbf{A}}(\mathcal{X}_{\mathbf{A}}). \tag{11}$$

Similarly, $\alpha_{\mathbf{B}}(\mathcal{X}_{\mathbf{B}}) = \alpha(\mathcal{X}_{\mathbf{A}\cup\mathbf{B}})$. Therefore, if $\mu(\mathcal{X}_{\mathbf{A}}) \neq \lambda(\mathcal{X}_{\mathbf{B}})$ there can be no measure α on $\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}$ satisfying $\alpha_{\mathbf{A}} = \mu$ and $\alpha_{\mathbf{B}} = \lambda$. In some situations, this problem is not too severe. Using the alternative definition, we express $\mu = \nu_1 H_1$ and $\lambda = \nu_2 H_2$. The consistency conditions translate to $H_1 = H_2$ and $\nu_1 = \nu_2$. If only the second condition fails, then it is still possible to find $H = H_1 \star H_2$. Employing the stick-breaking condition, we can generate random atoms from H. The problem lies in assigning weights to each atom. Fortunately, in density estimation, the value of the prior precision (ν) is typically small compared to

the sample size (n). If the estimate is robust to changes in ν, we may justifiably scale the base measures so that ν_1 and ν_2 are equal. In this case, it is only important that H_1 and H_2 are consistent. In other words, the base measures μ and λ only need to be *proportional* to each other over $\mathbf{A} \cap \mathbf{B}$.

There may be other situations in which scale *is* important. Unfortunately, as Equation 11 shows, we cannot find a suitable base measure for the prior that satisfies both μ and λ. Without a suitable prior, there can be no suitable posterior. If the goal is to estimate a distribution and there is genuine concern about the precision of the prior estimate, then both conditions must be satisfied. That is, μ and λ must be consistent. Equivalently, H_1 and H_2 must be consistent and ν_1 must be equal to ν_1. Cases in which one or both conditions fail are explored more fully in Appendix A.

Subsequently, we assume that both consistency conditions are satisfied. This leads to a natural extension of the previous work. We have equated consistency of base measures with consistency of probability measures. Thus, we generalize Markov combinations to include consistent finite measures by scaling them to probability measures, finding the Markov combination, and rescaling the measures.

Definition 12 (Markov Combination of Finite Measures**).** Let μ be a finite measure on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$. Let λ be a finite measure on $(\mathcal{X}_\mathbf{B}, \mathcal{F}_\mathbf{B})$ that is consistent with μ. The Markov combination of μ and λ is denoted $\mu \star \lambda$, where

$$\mu \star \lambda = \mu(\mathcal{X}_\mathbf{A}) \cdot [\overline{\mu} \star \overline{\lambda}] = \lambda(\mathcal{X}_\mathbf{B}) \cdot [\overline{\mu} \star \overline{\lambda}], \tag{12}$$

where $[\overline{\mu} \star \overline{\lambda}]$ is the almost-everywhere unique probability distribution satisfying Theorem 9.

This definition is a generalization of Definition 10 for probability measures. Note that the Markov combination defined in this way is unique almost everywhere, since $[\overline{\mu} \star \overline{\lambda}]$ is unique almost everywhere.

It is easy to show that the $\bar{\cdot}$ and $\star$ operations commute (with respect to composition).

Theorem 13. *If μ and λ are consistent measures, then $\overline{\mu \star \lambda} = \overline{\mu} \star \overline{\lambda}$.*

Proof.

$$\overline{\mu \star \lambda} = \frac{[\mu \star \lambda]}{[\mu \star \lambda](\mathcal{X}_{\mathbf{A} \cup \mathbf{B}})} \tag{13}$$

$$= \frac{\mu(\mathcal{X}_\mathbf{A}) \cdot [\overline{\mu} \star \overline{\lambda}]}{\mu(\mathcal{X}_\mathbf{A}) \cdot [\overline{\mu} \star \overline{\lambda}](\mathcal{X}_{\mathbf{A} \cup \mathbf{B}})} \tag{14}$$

$$= \overline{\mu} \star \overline{\lambda}. \tag{15}$$

□

Writing the base measures in the alternative notation, set $\mu = \nu H_1$ and $\lambda = \nu H_2$. Theorem 13 states that $\overline{\mu \star \lambda} = H_1 \star H_2$. Therefore, the Markov combination of νH_1 and νH_2 can be written $\nu(H_1 \star H_2)$.

4.3. Constructing the hyper Dirichlet Process

Suppose that $\mathcal{G}$ has two cliques and we desire to use a Dirichlet process prior for each one, with specific base measures. If we want to find a suitable Dirichlet process for the entire graph, we use the Markov combination of the two marginal base measures. As a result, the prior will have the desired marginal laws.

Theorem 14. *Let H_1 be a distribution on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$. Let H_2 be a distribution on $(\mathcal{X}_\mathbf{A}, \mathcal{F}_\mathbf{A})$ that is consistent with H_1. Set $H = H_1 \star H_2$. Let $Q \sim DP(\nu H_1)$, $R \sim DP(\nu H_2)$, and $\theta \sim DP(\nu H)$ be random probability measures. The following are true:*

(i) $\theta_\mathbf{A} \stackrel{d}{=} Q$

(ii) $\theta_\mathbf{B} \stackrel{d}{=} R$

Proof. The proposition follows from Theorem 5. □

As expected, the distribution of θ has the desired Dirichlet processes for marginals. Since this distribution is a Dirichlet process, previous results apply to this construction. Most importantly, we know that the prior law is a stick-breaking prior. Analogous parametric constructions, such as the hyper inverse Wishart family, have the desirable property of being *hyper Markov*.

Definition 15 (Hyper Markov). Consider an undirected graph, $\mathcal{G}$. Let $\theta \sim \mathcal{L}$ be a random probability measure over $\mathcal{X}$. We say that $\mathcal{L}$ is (weak) hyper Markov on $\mathcal{G}$ if $\mathcal{L}$ is concentrated on $\mathscr{M}(\mathcal{G})$, and $\theta_\mathbf{A} \perp\!\!\!\perp \theta_\mathbf{B} | \theta_\mathbf{C}$ whenever $\mathbf{C}$ separates $\mathbf{A}$ and $\mathbf{B}$.

The hyper Markov property is desirable because it ensures that the random parameters will yield a distribution for the data that lies within the graphical family of interest. Furthermore, the hyper Markov property ensures that the random parameters also have a graphical distribution, which is computationally convenient, as discussed earlier. Therefore, we would like to know when the construction in Theorem 5 us hyper Markov, at which point we can feel justified in calling it a *hyper Dirichlet process*.

Let $\mathcal{L} = DP(\nu H)$ be a Dirichlet process law. Let $\mathcal{G}$ be any graph consisting of two cliques, $\mathbf{A}$ and $\mathbf{B}$, with separator $\mathbf{C}$. Using the stick-breaking construction, let $\vec{w} = (w_i : i \in \mathbb{N})$ be the random weights and $\vec{Z} = (Z_i : i \in \mathbb{N})$ be the atoms, which are iid observations from H. We use $Z_{i\mathbf{\Gamma}}$ to denote the components of Z_i belonging to a set $\mathbf{\Gamma}$. For example, the marginal of θ over $\mathbf{A}$ is $\theta_\mathbf{A} = \sum_{i \in N} w_i \delta_{Z_{i\mathbf{A}}}$.

Obviously, one condition for hyper Markovity is that H is a Markov measure. If H is not Markov, then for $Z_i \sim H$, it is not true that $Z_{i\mathbf{A}} \perp\!\!\!\perp Z_{i\mathbf{B}} | Z_{i\mathbf{C}}$. As a result, $\theta_\mathbf{A} \not\perp\!\!\!\perp \theta_\mathbf{B} | \theta_\mathbf{C}$. This condition is necessary, but not sufficient. In addition to the location of the random atoms, $\theta_\mathbf{B}$ contains information about the distribution of weights at each atom. We must ensure that $\theta_\mathbf{C}$ contains the information as well. To see this, consider an example for which $H_\mathbf{C}$ is a point mass. For $H \in \mathscr{M}(\mathcal{G})$, this implies that $H_\mathbf{A} \perp\!\!\!\perp H_\mathbf{B}$. Further suppose that $H_\mathbf{A}$ and $H_\mathbf{B}$ are *not* point masses. In this case, $\theta_\mathbf{B}$ implies certain constraints on

$\vec{w}$. For example, if each $Z_{i\mathbf{B}}$ is distinct, then the mass at each atom determines the random weights modulo permutation. Therefore, the second condition for $\mathcal{L}$ to be hyper Markov is that $\theta_{\mathbf{B}}$ contains no information about $\vec{w}$ that is not contained by $\theta_{\mathbf{C}}$. To begin, we use the condition expressed in the next theorem. The condition is sufficient, but more restrictive than necessary.

Theorem 16. *Let H be a base measure on $\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}$. Let $\mathbf{C} = \mathbf{A} \cap \mathbf{B}$. Set $\mathcal{L} = DP(\nu H)$ for some $\nu > 0$. Then $\mathcal{L}$ is hyper Markov on $\mathcal{X}_{\mathbf{A}\cup\mathbf{B}}$ if the following conditions hold:*

1. *H is a Markov measure*
2. *Refinement Condition:*

$$Z_{i\mathbf{C}} = Z_{j\mathbf{C}} \Rightarrow Z_{i\mathbf{B}} = Z_{j\mathbf{B}} \quad \text{a.s.}[H].$$

Proof. Define $A' = \mathbf{A} \setminus \mathbf{C}$ and $B' = \mathbf{B} \setminus \mathbf{C}$. Note that $\mathbf{B} = \mathbf{C} \cup \mathbf{B}'$, so that $Z_{i\mathbf{B}} = Z_{j\mathbf{B}} \Rightarrow (Z_{i\mathbf{C}}, Z_{i\mathbf{B}'}) = (Z_{j\mathbf{C}}, Z_{i\mathbf{B}'})$. In other words, the refinement condition can be expressed equivalently as an "if and only if" statement:

$$Z_{i\mathbf{C}} = Z_{j\mathbf{C}} \Longleftrightarrow Z_{i\mathbf{B}} = Z_{j\mathbf{B}} \quad \text{a.s.}[H].$$

Consider $\theta \sim \mathcal{L}$. The hyper Markov property has two conditions:

1. $\mathbb{P}(\theta \in \mathscr{M}(\mathcal{G})) = 1$, and
2. $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$.

The first condition follows from the refinement condition. Let $x = (x_{\mathbf{A}'}, x_{\mathbf{C}}, x_{\mathbf{B}'})$ be any point in $\mathcal{X}$ such that $\theta_{\mathbf{B}}(x_{\mathbf{B}}) > 0$. That is, there exists some i such that $Z_{i\mathbf{B}} = x_{\mathbf{B}}$. By the refinement condition, $Z_{j\mathbf{C}} = Z_{i\mathbf{C}} = x_{\mathbf{C}}$ if and only if $Z_{j\mathbf{B}} = Z_{i\mathbf{B}} = x_{\mathbf{B}}$. Hence, $\{j : Z_{j\mathbf{C}} = x_{\mathbf{C}}\} = \{j : Z_{j\mathbf{B}} = x_{\mathbf{B}}\}$. Using the stick-breaking representation, we write the distribution of $X_{\mathbf{A}} | X_{\mathbf{B}}$.

$$\theta_{\mathbf{A}|\mathbf{B}}(x_{\mathbf{A}} | X_{\mathbf{B}} = x_{\mathbf{B}}) = \frac{\sum_{i:Z_{i\mathbf{B}}=x_{\mathbf{B}}} w_i 1_{\{Z_{i\mathbf{A}}=x_{\mathbf{A}}\}}}{\sum_{i:Z_{i\mathbf{B}}=x_{\mathbf{B}}} w_i} \tag{16}$$

$$= \frac{\sum_{i:Z_{i\mathbf{C}}=x_{\mathbf{C}}} w_i 1_{\{Z_{i\mathbf{A}}=x_{\mathbf{A}}\}}}{\sum_{i:Z_{i\mathbf{C}}=x_{\mathbf{C}}} w_i} \tag{17}$$

$$= \theta_{\mathbf{A}|\mathbf{C}}(x_{\mathbf{A}} | X_{\mathbf{C}} = x_{\mathbf{c}}). \tag{18}$$

Therefore, $\theta \in \mathscr{M}(\mathcal{G})$.

It remains to show that $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$. We begin by writing the marginals of θ using the stick-breaking representation. Let Γ be any subset of $\mathbf{V}$.

$$\theta_{\mathbf{\Gamma}} = \sum_{i\in\mathbb{N}} w_i \delta_{Z_{i\mathbf{\Gamma}}}. \tag{19}$$

Let $\vec{Z}^*_{\mathbf{\Gamma}} = \{Z_{i\mathbf{\Gamma}}\}$ be the set of unique occurrences among the random atoms. We refer to an element of this set using an arbitrary index, $Z^*_{i\mathbf{\Gamma}}$. Let $m^*_{i\mathbf{\Gamma}}$ be the total mass at that atom;

$$m^*_{i\mathbf{\Gamma}} = \sum_{j:Z_{j\mathbf{\Gamma}}=Z^*_{i\mathbf{\Gamma}}} w_j = \theta_{\mathbf{\Gamma}}(Z^*_{i\mathbf{\Gamma}}). \tag{20}$$

Note that $\vec{Z}^*_{\mathbf{\Gamma}}$ is the support of the $\theta_{\mathbf{\Gamma}}$, and $\vec{m}^*_{\mathbf{\Gamma}}$ is the mass at each point in the support. Thus, there is a bijection between the measure $\theta_{\mathbf{\Gamma}}$ and the set $\{\vec{Z}^*_{\mathbf{\Gamma}}, \vec{m}^*_{\mathbf{\Gamma}}\}$. That is to say, both are completely identified if at least one is known. The immediate result is that conditioning on one (or both) is equivalent to conditioning on the other (or one of them).

Continue by partitioning the support into two sets. Define $\vec{Z}^+_{\mathbf{\Gamma}} = \{Z^*_{i\mathbf{\Gamma}} : H_{\mathbf{\Gamma}}(Z^*_{i\mathbf{\Gamma}}) > 0\}$ and $\vec{Z}^0_{\mathbf{\Gamma}} = \{Z^*_{i\mathbf{\Gamma}} : H_{\mathbf{\Gamma}}(Z^*_{i\mathbf{\Gamma}}) = 0\} = \vec{Z}^*_{\mathbf{\Gamma}} \setminus \vec{Z}^+_{\mathbf{\Gamma}}$. In other words $\vec{Z}^+_{\mathbf{\Gamma}}$ is the set of support points with strictly positive mass under H_Γ and $\vec{Z}^0_{\mathbf{\Gamma}}$ is the set of points that are in the support but have probability zero under H_Γ. Again, we specify a particular element in either set with an arbitrary index, e.g. $Z^+_{i\mathbf{\Gamma}}$. Partition $\vec{m}^*_{\mathbf{\Gamma}}$ in the same way. This yields,

$$\vec{m}^+_{\mathbf{\Gamma}} = \{m^*_{i\mathbf{\Gamma}} : H_{\mathbf{\Gamma}}(Z^*_{i\mathbf{\Gamma}}) > 0\} = \{m^*_{i\mathbf{\Gamma}} : Z^*_{i\Gamma} \in \vec{Z}^+_{\mathbf{\Gamma}}\} = \{\theta_{\mathbf{\Gamma}}(Z^+_{i\mathbf{\Gamma}})\}. \tag{21}$$

We stipulate that the index is consistent with $\vec{Z}^+_{\mathbf{\Gamma}}$ so that $m^+_{i\mathbf{\Gamma}} = \theta_{\mathbf{\Gamma}}(Z^+_{i\mathbf{\Gamma}})$. Denote the other set in this partition by $\vec{m}^0_{\mathbf{\Gamma}} = \vec{m}^*_{\mathbf{\Gamma}} \setminus \vec{m}^+_{\mathbf{\Gamma}}$, where $m^0_{i\mathbf{\Gamma}} = \theta_{\mathbf{\Gamma}}(Z^0_{i\mathbf{\Gamma}})$. Separate the sum in Equation 19 using this partition.

$$\theta_{\mathbf{\Gamma}} = \sum_{i=1}^{N^+_{\mathbf{\Gamma}}} m^+_{i\mathbf{\Gamma}} \delta_{Z^+_{i\mathbf{\Gamma}}} + \sum_{i=1}^{N^0_{\mathbf{\Gamma}}} m^0_{i\mathbf{\Gamma}} \delta_{Z^0_{i\mathbf{\Gamma}}} \ , \tag{22}$$

where $N^\cdot_{\mathbf{\Gamma}} = |\vec{Z}^\cdot_{\mathbf{\Gamma}}|$. Note that $Z^+_{\mathbf{\Gamma}}$, has a degenerate distribution. If $H_{\mathbf{\Gamma}}(\gamma) > 0$, then with probability 1, γ will occur infinitely often in $\vec{Z}_{\mathbf{\Gamma}}$. Therefore, $\vec{Z}^+_{\mathbf{\Gamma}} = \{z_\Gamma : H_{\mathbf{\Gamma}}(z_\Gamma) > 0\}$ almost surely. Since H is known, the sets of summation in Equation 22 are fully identified by $\{\vec{Z}^*_{\mathbf{\Gamma}}, \vec{m}^*_{\mathbf{\Gamma}}\}$. It follows that conditioning on $\theta_{\mathbf{\Gamma}}$ is equivalent to conditioning on the quartet $\{\vec{Z}^+_{\mathbf{\Gamma}}, \vec{m}^+_{\mathbf{\Gamma}}, \vec{Z}^0_{\mathbf{\Gamma}}, \vec{m}^0_{\mathbf{\Gamma}}\}$. We will now show that under the refinement condition, $\vec{Z}^+_{\mathbf{B}}, \vec{m}^+_{\mathbf{B}}$, and $\vec{m}^0_{\mathbf{B}}$ are fully identified from $\theta_{\mathbf{C}}$. With that fact, showing $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$ is equivalent to showing $\theta_{\mathbf{A}} \perp\!\!\!\perp \vec{Z}^0_{\mathbf{B}} | \theta_{\mathbf{C}}$.

By the refinement condition,

$$m^+_{i\mathbf{C}} = \sum_{j:Z_{j\mathbf{C}}=Z^+_{i\mathbf{C}}} w_j = \sum_{j:Z_{j\mathbf{B}}=Z^+_{i\mathbf{B}}} w_j = m^+_{i\mathbf{B}}. \tag{23}$$

A similar equation shows $m^0_{i\mathbf{C}} = m^0_{i\mathbf{B}}$. Therefore, $(\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}) = (\vec{m}^+_{\mathbf{B}}, \vec{m}^0_{\mathbf{B}})$.

We now show that $\vec{Z}^+_{\mathbf{B}}$ is a function of $\vec{Z}^+_{\mathbf{C}}$. This fact ensures that $\vec{Z}^+_{\mathbf{B}}$ is fully identified by $\vec{Z}^+_{\mathbf{C}}$ and therefore is conditionally independent of anything given $\vec{Z}^+_{\mathbf{C}}$. One consequence of the refinement condition is that if $H_{\mathbf{C}}(c) > 0$, then

there exists $B(c)$ such that $H_{\mathbf{B}|\mathbf{C}}(B(c)|c) = 1$. This follows from a simple proof by contradiction. If $H_{\mathbf{B}|\mathbf{C}}(\cdot|c)$ is not a point distribution, than either every point has probability 0 (as in a continuous distribution), or there is some point with positive probability strictly less than 1. We will see that neither of these can be true and conclude the conditional is indeed a point distribution.

Suppose $H_{\mathbf{B}|\mathbf{C}}(b|c)$ has measure zero for every point $b \in \mathcal{X}_B$. With probability $H_{\mathbf{C}}(c)^2 > 0$, the event $Z_{1\mathbf{C}} = Z_{2\mathbf{C}}$ will occur. However, $Z_{1\mathbf{B}} \neq Z_{2\mathbf{B}}$ almost surely. Therefore, the refinement condition fails with probability at least $H_{\mathbf{C}}(c)^2 > 0$. Now suppose there exists b such that $0 < H_{\mathbf{B}|\mathbf{C}}(b|c) < 1$. Then with probability $H_{\mathbf{C}}(c)^2 H_{\mathbf{B}|\mathbf{C}}(b|c)(1 - H_{\mathbf{B}|\mathbf{C}}(b|c)) > 0$, the events $Z_{1\mathbf{C}} = c = Z_{2\mathbf{C}}$ and $Z_{1\mathbf{B}} = b \neq Z_{2\mathbf{B}}$ will occur. Thus, the refinement condition fails with positive probability. By these two contradictions, we see that $H_{\mathbf{B}|\mathbf{C}}(\cdot|c)$ must be a point distribution if $H_{\mathbf{C}}(c) > 0$. We denote the point of concentration by $B(c)$. Clearly, $c \in \vec{Z}^+_{\mathbf{C}}$ implies that $B(c) \in \vec{Z}^+_{\mathbf{B}}$. Furthermore, every element of $\vec{Z}^+_{\mathbf{B}} = B(c)$ for some $c \in \vec{Z}^+_{\mathbf{C}}$. This follows from the fact that $\mathbf{C} \subseteq \mathbf{B}$, so $H_{\mathbf{C}}(Z_{i\mathbf{C}}) = 0 \Rightarrow H_{\mathbf{B}}(Z_{i\mathbf{C}}, Z_{i\mathbf{B}'}) = 0$. Therefore, $\vec{Z}^+_{\mathbf{B}} = g(\vec{Z}^+_{\mathbf{C}}) = \{B(c) : c \in \vec{Z}^+_{\mathbf{C}}\}$ almost surely.

We have shown that conditioning on $\theta_{\cdot}$ is equivalent to conditioning on $\{\vec{Z}^+_{\cdot}, \vec{m}^+_{\cdot}, \vec{Z}^0_{\cdot}, \vec{m}^0_{\cdot}\}$. Furthermore, we have that $(\vec{Z}^+_{\mathbf{B}}, \vec{m}^+_{\mathbf{B}}, \vec{m}^0_{\mathbf{B}}) = (g(\vec{Z}^+_{\mathbf{C}}), \vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}})$. This provides an equivalent condition for the independence property that we want to show. That is, $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}}|\theta_{\mathbf{C}}$ if and only if $\theta_{\mathbf{A}} \perp\!\!\!\perp \vec{Z}^0_{\mathbf{B}}|\{\vec{Z}^+_{\mathbf{C}}, \vec{m}^+_{\mathbf{C}}, \vec{Z}^0_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}\}$. The remainder of this proof will show that the second property holds under the conditions of the theorem.

Begin by partitioning the atoms and weights as follows. Let $\hat{Z} = \{Z_i : Z_{i\mathbf{C}} \in \vec{Z}^+_{\mathbf{C}}\}$, and $\tilde{Z} = \{Z_i : Z_{i\mathbf{C}} \in \vec{Z}^0_{\mathbf{C}}\}$. Let $\hat{w} = \{w_i : Z_{i\mathbf{C}} \in \vec{Z}^+_{\mathbf{C}}\}$, and $\tilde{w} = \{w_i : Z_{i\mathbf{C}} \in \vec{Z}^0_{\mathbf{C}}\}$. As usual, for $\mathbf{\Gamma} \subseteq \mathbf{V}$, let $\hat{Z}_{\mathbf{\Gamma}}$ and $\tilde{Z}_{\mathbf{\Gamma}}$ denote that the elements are the components in $\mathbf{\Gamma}$. This partition is similar to, but different than, the partition defined earlier. $(\hat{Z}_{\mathbf{\Gamma}}, \tilde{Z}_{\mathbf{\Gamma}})$ depends on $H_{\mathbf{C}}$, whereas $(Z^+_{\mathbf{\Gamma}}, Z^0_{\mathbf{\Gamma}})$ depends on $H_{\mathbf{\Gamma}}$. The goal, as above is to rewrite $Z_{\mathbf{A}}$ by partitioning it in a way that preserves the conditional independence structure. This structure is preserved if the partioning function is non-random. In other words, the atoms must be partitioned based on on a known event. When conditioning on $\theta_{\mathbf{B}}$, $\theta_{\mathbf{B}}$ and $\theta_{\mathbf{C}}$ are known, but $\theta_{\mathbf{A}}$ is unknown. Therefore, $(\hat{Z}_{\mathbf{A}}, \tilde{Z}_{\mathbf{A}})$ provides an observable partition of $\vec{Z}_{\mathbf{A}}$.

$$\theta_{\mathbf{A}} = \sum_{i=1}^{N^+_{\mathbf{C}}} \hat{w}_i \hat{Z}_{i\mathbf{A}} + \sum_{i=1}^{N^0_{\mathbf{C}}} \tilde{w}_i \tilde{Z}_{i\mathbf{A}}. \tag{24}$$

Note that $\tilde{Z}_{\mathbf{C}}$ is equivalent to $\vec{Z}^0_{\mathbf{C}}$ by definition, and $\tilde{Z}_{\mathbf{B}} = \vec{Z}^0_{\mathbf{B}}$ by the refinement condition. We proceed by showing that $\tilde{w}, \tilde{Z}_{\mathbf{A}}, \hat{w}$, and $\hat{Z}_{\mathbf{A}}$ are jointly independent of $\vec{Z}^0_{\mathbf{B}}$ given $\{\vec{Z}^+_{\mathbf{C}}, \vec{m}^+_{\mathbf{C}}, \vec{Z}^0_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}\}$. We can express $\vec{m}^+_{\mathbf{C}}$ as a function of $\hat{w}, \hat{Z}$, and $\vec{Z}^+_{\mathbf{C}}$, where

$$m^+_{i\mathbf{C}} \overset{a.s.}{=} \sum_{j:\hat{Z}_j = Z^+_{i\mathbf{C}}} \hat{w}_j. \tag{25}$$

Furthermore, we have noted that $\vec{m}^0_{\mathbf{C}} = \tilde{w}$. By the stick-breaking construction, $\tilde{Z} \perp\!\!\!\perp (\tilde{w}, \hat{w}, \hat{Z})$. Since $\vec{Z}^+_{\mathbf{C}}$ is known almost surely, it can be included in the independence property.

$$\tilde{Z} \perp\!\!\!\perp (\vec{Z}^+_{\mathbf{C}}, \hat{Z}, \hat{w}, \tilde{w}). \tag{26}$$

$\tilde{Z}$ is also independent of any function of the RHS of Equation 26. In particular,

$$\tilde{Z} \perp\!\!\!\perp (\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}, \vec{Z}^+_{\mathbf{C}}, \hat{Z}_{\mathbf{A}}, \hat{w}, \tilde{w}). \tag{27}$$

Repeating this argument on the LHS of Equation 26, we conclude

$$(\vec{Z}^0_{\mathbf{B}}, \tilde{Z}_{\mathbf{A}}, \vec{Z}^0_{\mathbf{C}}) \perp\!\!\!\perp (\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}, \vec{Z}^+_{\mathbf{C}}, \hat{Z}_{\mathbf{A}}, \hat{w}, \tilde{w}). \tag{28}$$

Since $\tilde{Z}_i \sim H \in \mathscr{M}(\mathcal{G})$, we can write $\vec{Z}^0_{\mathbf{B}} \perp\!\!\!\perp \tilde{Z}_{\mathbf{A}} | \vec{Z}^0_{\mathbf{C}}$. Since all three of these are jointly independent of $(\hat{Z}_{\mathbf{A}}, \hat{w})$ and $(\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}, \vec{Z}^+_{\mathbf{C}}))$, it follows that

$$\vec{Z}^0_{\mathbf{B}} \perp\!\!\!\perp (\hat{Z}_{\mathbf{A}}, \hat{w}, \tilde{Z}_{\mathbf{A}}, \tilde{w}) | (\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}, \vec{Z}^+_{\mathbf{C}}, \vec{Z}^0_{\mathbf{C}}). \tag{29}$$

Recall from Equation 24 that $\theta_{\mathbf{A}}$ is a function of $(\hat{Z}_{\mathbf{A}}, \hat{w}, \tilde{Z}_{\mathbf{A}}, \tilde{w})$. It follows that,

$$\vec{Z}^0_{\mathbf{B}} \perp\!\!\!\perp \theta_{\mathbf{A}} | (\vec{m}^+_{\mathbf{C}}, \vec{m}^0_{\mathbf{C}}, \vec{Z}^+_{\mathbf{C}}, \vec{Z}^0_{\mathbf{C}}). \tag{30}$$

Hence, by the above argument, it follows that $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$. We conclude that $\mathcal{L}$ is hyper Markov. □

Theorem 16 provides sufficient conditions for a Dirichlet process to be hyper Markov. Thus, under those conditions, we may safely call the Dirichlet process a *hyper* Dirichlet process. When H satisfies the refinement condition, we will say that $\mathbf{B}$ is a refinement of $\mathbf{C}$ under sampling almost surely under measure H. It is a refinement in the following sense. Let $X_1, X_2, \ldots$ be an infinite iid sample from H. Form a partition of the natural numbers such that i and j are elements of the same set if and only if $X_{i\mathbf{C}} = X_{j\mathbf{C}}$. Call this partition $X(\mathbf{C})$. Define $X(\mathbf{B})$ by analogy. Under the refinement condition, $X(\mathbf{B})$ is almost surely a refinement of $X(\mathbf{C})$. We denote this relationship by $X(\mathbf{B}) \preceq X(\mathbf{C})$ $a.s.[H]$, omitting H if the measure is contextually evident.

The refinement condition, as stated in Theorem 16 is sufficient, but it is stronger than necessary. By symmetry of conditional independence, $\theta_{\mathbf{B}} \perp\!\!\!\perp \theta_{\mathbf{A}} | \theta_{\mathbf{C}}$, even though no refinement condition is needed between $\mathbf{C}$ and $\mathbf{A}$. It may be necessary that at least one of the two refinements is present, but this has not been explored.

The hyper Dirichlet process defined on two cliques is an example of a hyper Markov combination, which is the analog of Markov combinations for prior laws. Consider two laws: $\mathcal{Q}$ for $\theta_{\mathbf{A}}$ and $\mathcal{R}$ for $\theta_{\mathbf{B}}$. We say that $\mathcal{Q}$ and $\mathcal{R}$ are *hyperconsistent* if the marginal laws for $\theta_{\mathbf{A}\cup\mathbf{B}}$ are equal. Under this condition, Dawid and Lauritzen (1993) show that there is a unique hyper Markov law $\mathcal{L}$ such that $\mathcal{L}_{\mathbf{A}} = \mathcal{Q}$, $\mathcal{L}_{\mathbf{B}} = \mathcal{R}$. This is called the hyper Markov combination and is denoted $\mathcal{L} = \mathcal{Q} \odot \mathcal{R}$.

As with Markov combinations, hyper Markov combinations are easily generalized to multiple cliques. Let $\mathcal{G}$ be a graph with perfectly ordered cliques $(\mathbf{C}_1, \ldots, \mathbf{C}_k)$. Suppose $\mathbf{C}_i$ is imbued with a prior law $\mathcal{Q}_i$ and that the priors are all pairwise hyperconsistent. Let $\mathcal{L}_1 = \mathcal{Q}_1$ and $\mathcal{L}_i = \mathcal{L}_{i-1} \odot \mathcal{Q}_i$ for $i \geq 2$. Then $\mathcal{L} = \mathcal{L}_k$ is the unique hyper Markov prior satisfying $\mathcal{L}_{\mathbf{C}_i} = \mathcal{Q}_i$. We call $\mathcal{L}$ the hyper Markov combination of $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$. In general we may write $\odot(\mathcal{Q}_1, \ldots, \mathcal{Q}_k)$ with the understanding that the cliques are perfectly ordered and $\mathcal{Q}_1, \ldots, \mathcal{Q}_k$ are pairwise consistent.

The next definition generalizes the hyper Dirichlet Process to three or more cliques.

Definition 17 (Hyper Dirichlet Process**).** Let $\mathcal{G}$ be a graph with a perfect ordering of cliques $\mathbf{C}_1, \ldots, \mathbf{C}_k$. Suppose that the i^{th} clique has marginal distribution H_i and that the marginals are pairwise consistent. Let $H = \star(H_1, \ldots, H_k)$. Further suppose that $\mathbf{C}_j$ or $\mathbf{H}_j$ is a refinement of $\mathbf{S}_j$ under sampling almost surely under H, where $\mathbf{H}_i$ is the i^{th} history and $\mathbf{S}_i$ is the i^{th} separator. Then

$$\text{HDP}(\nu, S_1, \ldots, S_k) = \text{DP}(\nu S) \tag{31}$$

is a *hyper Dirichlet process* prior.

The hyper Dirichlet process defined in this way is guaranteed to be hyper Markov. In fact, it is the unique hyper Markov combination of the marginal Dirichlet processes. Suppose $\mathcal{L} = \text{HDP}(\nu, H_1, \ldots, H_k)$. By Theorem 7, $\mathcal{L}_{\mathbf{C}_i} = \text{DP}(\nu H_i)$ for $i \geq 2$. Furthermore, it follows from the refinement conditions and Theorem 16 that $\mathcal{L}_{\mathbf{H}_{i-1} \cup \mathbf{C}_i} = \text{DP}(\nu H_{\mathbf{H}_{i-1}}) \odot \text{DP}(\nu H_i)$. Hence, Theorem 3.9 in Dawid and Lauritzen (1993) states that $\mathcal{L}$ is the almost-everywhere unique hyper Markov law such that $\mathcal{L}_{\mathbf{C}_i} = \text{DP}(\nu H_i)$, which by definition is called the hyper Markov combination. In other words,

$$\begin{aligned} \text{HDP}(\nu, H_1, \ldots, H_k) &= \text{DP}(\nu \star (H_1, \ldots, H_k)) & (32) \\ &= \odot(\text{DP}(\nu H_1), \ldots, \text{DP}(\nu H_k)). & (33) \end{aligned}$$

4.4. Properties of the hyper Dirichlet process

By definition 17, the hyper Dirichlet process is a Dirichlet process with a specialized base measure. Thus, *any result proved for Dirichlet processes also applies to hyper Dirichlet processes.* For example, Theroem 7 states that the posterior law will also be a Dirichlet process with an easily updated base measure. The next theorem strengthens this result by showing that the posterior will also be a hyper Dirichlet process.

Theorem 18 (Posterior Hyper Dirichlet Process**).** *Suppose $\mathcal{G}$ is a graph with a perfect ordering of cliques $\mathbf{C}_1, \ldots, \mathbf{C}_k$. Let $\mathcal{L} = \text{HDP}(\nu, H_1, \ldots, H_k)$ be a hyper Dirichlet process. Given $\theta \sim \mathcal{L}$, let $X_1, \ldots, X_n$ be an iid sample from θ. Denote the marginal value of X_j over the i^{th} clique by X_{ji}. Then,*

$$\theta | X_1, \ldots, X_n \sim \text{HDP}(\nu', H_1', \ldots, H_k'), \tag{34}$$

where $H_i' = (\nu + n)^{-1}(\nu H + \sum_j \delta_{X_{ji}})$

Proof. By Theorem 7 above, the posterior is $\mathrm{DP}(\nu' H')$ where $H' = (\nu+n)^{-1}(\nu H + \sum_j \delta_{X_j})$. By Corollary 5.2 in Dawid and Lauritzen (1993), this posterior must be hyper Markov because the prior is hyper Markov. This in part implies that H' is Markov. Since it has the correct marginals, H' must be the Markov combination of the marginal updates: $H' = \star(H'_1, \ldots, H'_k)$. The result now holds by Definition 17. □

5. Applications

According to our definition of a hyper Dirichlet process (Definition 17), a hyper Dirichlet process is a Dirichlet process with a base measure that satisfies certain properties. This is in line with various families of parametric hyper Markov distributions. For example, the hyper Normal is a Normal distribution with a constrained covariance matrix. The implication of this property is that a wide variety of results and applications for Dirichlet processes can be easily extended to situations which call for the *hyper* Dirichlet process.

While the refinement condition seems unduly restrictive at first glance, it allows one to use hyper Dirichlet processes in most areas that have benefited from Dirichlet processes. For the many applications that use a continuous base measure, the random atoms are distinct with probability one, so the refinement condition is trivial. Furthermore, Theorem 18 states that the posterior will also be hyper Markov. As a result, the hyper Dirichlet process is suitable for applications such as MCMC, in which one needs a posterior update that is also hyper Markov. In the following subsections, we study mixture models to illustrate how the hyper Dirichlet process provides a convenient extension of Dirichlet process theory.

5.1. Hyper Dirichlet mixtures

One place where the Dirichlet process shines is in mixture modeling. As an example application, we will see that the hyper Dirichlet process allows us to incorporate graphical structures into a mixture model. We begin by reviewing the non-hyper version of Dirichlet mixtures.

Suppose $D_n = X_1, \ldots, X_n$ are observations from some family of distributions parameterized by π. If we allow π_i to vary for each observation, then the result is a mixture of distributions. The number of parameters increases with n, which necessitates placing some prior on the distribution of π. If the prior is unknown, it can be modeled with a Dirichlet process. In general, a Dirichlet mixture is a hierarchical model expressed as

$$\theta \sim DP(\nu H) \tag{35}$$

$$\pi_1, \ldots, \pi_n | \theta \sim \theta \tag{36}$$

$$X_i | \pi_1, \ldots, \pi_n \sim f(X|\pi_i). \tag{37}$$

Combining both parts of Theorem 7, the conditional distribution of π_n given $\pi_1, \ldots, \pi_{n-1}$ is

$$H' = (\nu/\nu')H + \sum_{j=1}^{n-1}(1/\nu')\delta_{\pi_j}, \tag{38}$$

where $\nu' = \nu + n - 1$. Thus, with positive probability, π_n will be a previous value of π_i; otherwise, it is drawn from H. As a result, there will be $k \leq n$ unique values of π_i. This induces a latent class model in which each class is defined by a shared value of π_i. A key feature is that the number of latent classes is estimated rather than specified *a priori*. It is clear from Equation 38 that this estimate is influenced by ν. When ν is large, new values of π_i will often be drawn from H. Contrarily, when ν is small, π_i will more often be drawn from the previous values. Therefore, ν implies a prior distribution for the number of components.

Note that in the model specified above, the data are conditionally independent given the latent class assignments. Mathematically, $\perp\!\!\!\perp \{X_i\}|\{\pi_i\}$. Therefore, the predictive distribution of the next observation conditioned on the current sample and parameters satisfies

$$F(X_{n+1}|\pi, D_n) = \int F(X_{n+1}|\pi_{n+1})dF(\pi_{n+1}|\pi) = F(X_{n+1}|\pi), \tag{39}$$

where $\pi = \pi_1 \ldots \pi_n$. Bayesian density estimation is solved by using the prior to integrate out the unknown π. This integral is intractable, but it can be estimated using a Gibbs sampler for some models. Let $\pi^{(i)} = \pi \setminus \{\pi_i\}$ be the parameters other than π_i. The required conditional for π_i is

$$F(\pi_i|\pi^{(i)}, D_n) \propto \nu f_H(x_i) \cdot H_i(\pi_i) + \sum_{j=1, j\neq i}^{n} f(x_i|\pi_j) \cdot \delta_{\pi_j}(\pi_i), \tag{40}$$

where H_i is the posterior for $\pi_i|x_i$ and f_H is the marginal density for x_i when $\pi_i \sim H$. Note that by incorporating the data, the probability of drawing a previously seen π_j depends on $f(x_i|\pi_j)$. In the terminology of latent class models, x_i is more likely to be assigned the same class as x_j if it has a higher probability under the class' model. For example, if π_j is a location parameter and f is unimodal, data near each other are more likely to be assigned the same label compared to more distant data. Equation 40 reveals the two requirements necessary for using this type of Gibbs sampler. First, we must be able to calculate the marginal density of x_i. Secondly, we must be able to *sample* from the posteriors, $\{H_i\}$.

As an extenstion of this idea, suppose that the data exhibits some conditional independence structure specified by the graph $\mathcal{G}$. We need only update the model at the highest level by specifying a *hyper* Dirichlet process for θ:

$$\theta \sim HDP_{\mathcal{G}}(\nu H) \tag{41}$$

$$\pi_1, \ldots, \pi_n|\theta \sim \theta \tag{42}$$

$$X_i|\pi_1, \ldots, \pi_n \sim f(X|\pi_i). \tag{43}$$

This yields a hyper Markov structure in the following sense. Once again, we induce latent classes by assigning x_i and x_j the same class if $\pi_i = \pi_j$. Given its class label, each sample observation has a distribution which is Markov with respect to $\mathcal{G}$. Furthermore, the unique class parameters $\{\pi_1 \dots \pi_n\}$ are iid draws from a hyper Markov prior. Thus, by taking advantage of the hyper Dirichlet process, we are able to incorporate graphical knowledge into a new type of model. Furthermore, the hyper Markov structure is computationally convenient since calculations can be carried out individually over each clique and separator as we see in the following Gaussian example.

5.2. Example: hyper Dirichlet mixture of Gaussians

One example of the Dirichlet mixture model is the Dirichlet mixture of Gaussians, in which $\pi_i = (\mu_i, V_i)$ and $X_i|\pi_i \sim N(\mu_i, V_i)$. In this case, H is a measure over $(\mathcal{R}, M_p^+)$. Escobar and West (1995) describe a univariate mixture in which they define H to be a Normal $\times$ Inverse-Gamma distribution. The variance, V_i is Inverse-Gamma, where $V_i^{-1} \sim F(s/2, S/2)$, a Gamma distribution with shape $s/2$ and scale $S/2$. Conditional on the variance, the mean has Normal distribution, $N(\mu_i, \tau V_i)$. This prior is conjugate to the Normal distribution. The posterior for $(V_i|X_i)$ is Inverse-Gamma with $V_i^{-1} \sim G((1+s)/2, S_i/2)$, where $S_i = S + (x_i - m)^2/(1+\tau)$. The conditional posterior for $(\mu_i|V_i, x_i)$ is $N((m+\tau x_i)/(1+\tau), \tau V_i/(1+\tau))$. The marginal distribution of x_i is $T(s, m, M)$, the t-distribution with s degrees of freedom, non-centrality m, and scale $\sqrt{M}$, where $M = (t+\tau)S/s$. Thus, a Gibbs sampler is possible, taking advantage of the conditionals specified in Equation 40. The hyper Dirichlet process allows us to extend the Dirichlet mixture model to a multidimensional and graphical setting. Here, we restrict the base measure H to satisfy Theorem 16, so the Dirichlet process is now a hyper Dirichlet process. In particular, we restrict its support to $(\mathcal{R}, Q_\mathcal{G})$. Recall that $V \in Q_\mathcal{G}$ if $V_{ab}^{-1} = 0$ whenever there is no edge between a and b in $\mathcal{G}$. A relatively simple method is to update the $N \times IG$ model by replacing the Inverse-Gamma with its hyper Markov cousin, the hyper inverse Wishart distribution. In this model, a new mixture component is drawn as follows:

$$V_i \sim HIW_\mathcal{G}(d, D) \tag{44}$$

$$\mu_i|V_i \sim N_p(m, \tau V_i) \tag{45}$$

$$X_i|\mu_i, V_i \sim N_p(\mu_i, V_i), \tag{46}$$

where $HIW_\mathcal{G}(d, D)$ is the hyper inverse Wishart distribution with center D and d degrees of freedom. Note that the first two lines specify, H, a hyper Markov base measure for (μ_i, V_i). Since this measure is continuous, the Refinement Condition is satisfied trivially. Thus, the end result will be a *hyper Dirichlet mixture of Gaussians*.

In addition to being hyper Markov, H is also conjugate. Consider observing a random variable, X, where $X \sim N(\mu, V)$ and the parameters $(\mu, V) \sim H$

are unknown. The posterior distribution is $f(\mu, V|X) = f(V|X)f(\mu|V, X)$. By conditioning on V, the posterior calculation for μ reduces to the Normal-Normal conjugate model with known covariance. Thus, $(\mu|V, X)$ is Normal with mean $(\tau m + X)/(\tau + 1)$ and covariance $\tau V/(\tau + 1)$. Furthermore, $X|V$ is marginally normal with mean m and covariance $(1 + \tau)V$. This gives an expression for the marginal model, integrated over all μ.

$$\begin{aligned} V &\sim HIW_{\mathcal{G}}(d, D) \\ (X|V) &\sim N(m, (1 + \tau V)) \end{aligned}$$

Taking advantage of the hyper Markov structure, the posterior can be found by updating each clique and separator individually. Thus, the posterior distribution of V after observing X is

$$(V|X) \sim \frac{\prod_{c\in\mathcal{C}} dIW(V_c; d + 1, D_c + \Phi_c)}{\prod_{s\in\mathcal{S}} dIW(V_s; d + 1, D_s + \Phi_s)}, \tag{47}$$

where $\Phi_c = x_c x_c'$ is 1-sample covariance matrix for clique c. Therefore, the posterior distribution of $(V|X)$ is $HIW(d + 1, D + \Phi_{\mathcal{G}})$, where $\Phi_{\mathcal{G}}$ is the 1-sample covariance matrix under the graphical constraint.

Taking further advantage of the hyper Markov structure, we find the marginal distribution for X by integrating each clique and separator individually. My model leads to a new Markov distribution. The marginal distribution of X for the 1-sample problem is

$$X \sim \frac{\prod_{c\in\mathcal{C}} dT\left(X_c; d + 1 - |c|, m_c, \frac{\tau+1}{d+1-|c|} D_c\right)}{\prod_{s\in\mathcal{S}} dT\left(X_s; d + 1 - |s|, m_s, \frac{\tau+1}{d+1-|s|} D_s\right)}, \tag{48}$$

where $dT(x; d, m, D)$ represents the density evaluated at x of the multivariate t-distribution having d degrees of freedom, non-centrality parameter m, and scale parameter D, and $|c|$ is the number of elements in c. We call this the *hyper t-distribution* because it generalizes the multivariate-t in the same way that hyper inverse Wishart generalizes the inverse Wishart. The notation for the hyper t-distribution specified by Equation 48 is $HT(d + 1, m, (\tau + 1)D)$. Dawid and Lauritzen (1993) present what they call the *matrix T* distribution, which is a special case of this hyper t-distribution in which $m = 0$. This case is not general enough for my model. As a mixture of several distributions with different centers, we need to consider cases in which $m \neq 0$.

Following Equation 40 we can incorporate a Gibbs sampler to solve the Bayesian density estimation. The marginal density calculation is simply the product of marginal t-distributions. The posterior samples from hyper inverse Wishart distributions can follow the algorithm detailed by Carvalho et al. (2007).

6. Discussion

The hyper Dirichlet process has been defined by constraining the base measure of a Dirichlet process. Therefore, any result pertaining to Dirichlet process can be

reused in the hyper Markov case. We have shown one example of how the hyper Dirichlet process provides a convenient extension to Dirichlet process theory. As a final note, we point out that some of the results apply to other stick-breaking measures. Notably, Theorem 16 did not rely on the distribution of the random weights. Therefore, the same conditions imply that any stick-breaking measure is hyper Markov. That is, if the H is Markov and the refinement condition holds, then a stick-breaking prior whose atoms have distribution H is a hyper Markov law. Whether or not the posterior is also hyper Markov depends on how the measure is updated. For the Dirichlet process, the posterior update mechanism ensures a hyper Markov posterior as long as the prior is hyper Markov.

7. Acknowledgements

The author would like to thank Profs. S. Fienberg, A. Rinaldo, C. Schafer, and C. Shalizi for their guidance in developing this paper.

Appendix A: Working with non-consistent base measures

We constructed the hyper Dirichlet process by combining base measures on each clique into a base measure on the entire graph. As we have seen, the end result is simply another Dirichlet process. In other words, the hyper Dirichlet process is simply a Dirichlet process that is hyper Markov on the graph of interest. The benefit is that the elicitation of the base measure is simplified by only considering a subset of variables at a time. In the current paper, we considered the case in which the individual base measures are consistent (i.e. they agree about the intersection.) However, we may not be able to guarantee consistency, especially if component base measures are elicited from different experts or models. How can this be resolved?

Consider a two cliques, $\mathbf{A}$ and $\mathbf{B}$ with intersection $\mathbf{C}$. Let μ be a measure on $\mathcal{X}_{\mathbf{A}}$ and λ be a measure on $\mathcal{X}_{\mathbf{B}}$. The two measures are consistent if the marginals over $\mathcal{X}_{\mathbf{C}}$ are equal. In Section 4.2 we stated that this can be expressed as two simultaneous conditions: (1) the marginals must be proportional, and (2) the marginals must have the same scale.

Suppose only the first condition holds. Recall that $\mu(\mathcal{X}_{\mathbf{A}})$ represents the prior sample size. The interpretation is that there is more prior information about one clique than there is about the other clique. Equation 11 shows that there is no Markov Combination of μ and λ. That is, $\mu \star \lambda$ does not exist. Thus, there is no hyper Dirichlet process with those marginals. Fortunately, it is still possible to generate an "almost appropriate" random distribution. This is possible because we only need $\overline{\mu \star \lambda}$. By the commutative property, we can use $\overline{\mu} \star \overline{\lambda}$ instead. This is well-defined since the first property ensures that $\overline{\mu}$ and $\overline{\lambda}$ are consistent. On the other hand, the difference in scale must be resolved if a problem requires having a well-defined prior or posterior. The simplest way to achieve this is to scale one measure up or down to match the other. Additionally, any convex combination of $\mu\mathcal{X}_{\mathbf{A}}$ and $\lambda\mathcal{X}_{\mathbf{B}}$ could be a logical choice. The most conservative

choice would be to set $\mu(\mathcal{X}_{\mathbf{A}}) = \lambda(\mathcal{X}_{\mathbf{B}}) = \min\{\mu(\mathcal{X}_{\mathbf{A}}), \lambda(\mathcal{X}_{\mathbf{B}})\}$. If resources are sufficient, both scales could be used and the results compared. This would reveal how sensitive the outcome is to the scale of the base measures.

Now suppose that the first condition does not hold. The interpretation is that we have conflicting prior information. Once again, $\mu \star \lambda$ does not exist. Furthermore, $\overline{\mu} \star \overline{\lambda}$ does not exist either, so it is not possible to use the same method to generate random distributions. In order to find a base measure, one or both distributions must be changed. There are several natural ways to do this. Let $U \subseteq \mathcal{X}_{\mathbf{A}}, V \subseteq \mathcal{X}_{\mathbf{B}}, W \subseteq \mathcal{X}_{\mathbf{C}}$.

1. Choose one base measure, and complete the distribution via conditioning.

$$\alpha^{\mathbf{A}}(U \times V \times W) = \mu(U \times W)\lambda(V|W). \tag{49}$$
$$\alpha^{\mathbf{B}}(U \times V \times W) = \mu(U|W)\lambda(V \times W). \tag{50}$$

2. Calculate a weighted average.

$$\alpha^{w}(U \times V \times W) = \gamma\alpha^{\mathbf{A}} + (1-\gamma)\alpha^{\mathbf{B}}, \quad \gamma \in [0,1]. \tag{51}$$

   If there is no reason to choose one prior over the other, $\gamma = 1/2$ is appropriate. An interesting choice of γ is $\mu(\mathcal{X}_{\mathbf{A}})/(\mu(\mathcal{X}_{\mathbf{A}}) + \lambda(\mathcal{X}_{\mathbf{B}}))$. This gives more weight toward the prior with more information.
3. Minimize the summed KL-divergence. Let $\mu_{\mathbf{C}}$ and $\lambda_{\mathbf{C}}$ be the marginals over $\mathcal{X}_{\mathbf{C}}$.

$$\alpha_{\mathbf{C}} = \operatorname*{argmin}_{\nu} \int_{\mathcal{X}_{\mathbf{C}}} \overline{\mu}_{\mathbf{C}} \frac{\mu_{\mathbf{C}}}{\nu} + \int_{\mathcal{X}_{\mathbf{C}}} \overline{\lambda}_{\mathbf{C}} \frac{\lambda_{\mathbf{C}}}{\nu}. \tag{52}$$
$$\alpha(U \times V \times W) = \mu(U|W)\alpha_{\mathbf{C}}(W)\lambda(V|W). \tag{53}$$

More work is needed to test these candidate solutions to form good recommendations about their use.

References

Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika*, **83** 275–285.

Carvalho, C., Massam, H. and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, **94** 647–659. http://ftp.stat.duke.edu/WorkingPapers/05-03.html. MR2410014

Dawid, A.P. and Lauritzen, S.L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21** 1272–1317. MR1241267

Dempster, A. (1972). Covariance selection. *Biometrics*, **28** 157–175.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90** 196–212. MR2064941

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90** 577–588. MR1340510

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

Ghosh, J.K. and Ramamoorthi, R. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. In *Analysis of Censored Data* (H. Koul and J. Deshpande, eds.). 95–104. MR1483342

Giudici, P. and Green, P. (1999). Decomposable graphical gaussian model determination. *Biometrika*, **86** 785–801. MR1741977

Griffin, J.E. and Steel, M.F.J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, **101**. MR2268037

Holland, P.W. and Leinhardt, S. (1981). An exponential family of probability distributions. *Journal of the American Statistical Association*, **76** 33–50. MR0608176

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *The Annals of Statistics*, **29** 668–686. MR1865336

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics*, **35** 1278–1323. MR2341706

Liu, J. and Massam, H. (2006). The conjugate prior for discrete hierarchical log-linear models. `http://www.citebase.org/abstract?id=oai:arXiv.org:math/0609100`.

Marrelec, G. and Benali, H. (2006). Asymptotic Bayesian structure learning using graph supports for Gaussian graphical models. *Journal of Multivariate Analysis*, **97** 1451–1466. MR2256161

Pattison, P. and Wasserman, S. (1999). Logit models and logistic regression for social networks: II. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, **52** 169–193.

Pievatolo, A. and Rotondi, R. (2000). Analysing the interevent time distribution to identify seismicity phases: A Bayesian nonparametric approach to the multiple changepoint problem. *Applied Statistics*, **49** 543–562. MR1824558

Robins, G., Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks: III. valued relations. *Psychometrika*, **64** 371–394. MR1720089

Roverato, A. and Whittaker, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models. *Biometrika*, **85** 711–725. MR1665842

Schervish, M.J. (1995). *Theory of Statistics*. Springer-Verlag, New York. MR1354146

Sethuraman, J. (1994). A constructive definition of Dirichlet measures. *Statistica Sinica*.

Speed, T. and Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, **14** 138–150. MR0829559

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, **85** 204–212. MR1137368

Susarla, V. and Ryzin, J.V. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71** 897–902. MR0436445

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. *Psychometrika*, **61** 401–425. MR1424909