Thesis Proposal: Non-parametric Hyper Markov Priors

Daniel Heinz

September 25, 2008

Abstract

Markov distributions are used to describe multivariate data with conditional independence structure. Applications of Markov distributions arise in many fields including demography, flood prediction, and telecommunications. A hyper Markov law is a distribution over the space of all Markov distributions; such laws have been used as prior distributions for various types of graphical models. Dirichlet processes have also been used to specify priors in a non-parametric form. I have developed a family of non-parametric hyper Markov laws that I call hyper Dirichlet processes, which combine the separate ideas of hyper Markov laws and non-parametric prior processes. In my thesis, I propose to describe these distributions and their properties, and to apply them to specific problems. For example, I define a hyper Markov mixture of Gaussians and use it in the form of a hyper Markov prior to provide a non-parametric way to mix graphical Gaussian distributions.

1 Introduction

Markov distributions are multivariate measures that satisfy a specified set of conditional independence relations, often represented by an undirected graph. A measure is Markov with respect to a graph if two variables are conditionally independent whenever there is no edge between them in the graph. Markov distributions, or *Markov random fields*, have been used for a wide variety of problems, including demography (Sebastiani, 2003), flood prediction (Allcroft and Glasbey, 2003), and telecommunications (Zachary and Ziedins, 1999).

Dawid and Lauritzen (1993) extended the notion of Markov distributions for variables to *hyper Markov* distributions for parameters. In Bayesian statistics, one considers a random distribution, which therefore has its own distribution called the prior. A prior law over Markov measures is *hyper Markov* if the random marginal measures also satisfy the conditional independence structure. This is equivalent to requiring that the distribution of each variable is conditionally independent of the joint distribution of the other variables given the joint distribution of its neighbors. Importantly, hyper Markov priors reduce the size of the model space, so they are more efficient if the conditional independence structure is correct. When the hyper Markov prior is constrained to the space of Gaussian Markov distributions, the result is a graphical Gaussian model, an object of much study (Giudici and Green, 1999; Roverato, 2000; Carvalho et al., 2007; Letac and Massam, 2007; Banerjee et al., 2007), but inherently limited by the strong assumption of normality.

Non-parametric priors are used to avoid assumptions about the shapes of unknown distributions, whether univariate or multivariate. A popular example is the Dirichlet process (Ferguson, 1973), which Escobar and West 1995 use with mixtures of Gaussians.

Unfortunately, there has been little research in cases for which the variables exhibit an independence structure. This void is the focus of my research. In Heinz (submitted), I introduce the *hyper Dirichlet process*, which is a non-parametric hyper Markov prior. This is summarized in Section 7. This process combines the benefits of the non-parametric approach with the hyper Markov literature. Furthermore, I define the process in such a way that previous research about Dirichlet processes applies to the hyper case. As a result, many applications of the Dirichlet process can easily be generalized to graphical models. As

an example, I describe a new model, a hyper Dirichlet mixture of Gaussians. This model is a mixture of Gaussian components, each of which satisfies the given independence constraints.

The major goals of my research are to pursue my mixture model, which I can fit the model using a Gibbs sampler. I will determine good diagnostics to see when the sampler has converged. I will also expand this algorithm to incorporate learning about hyperparameters and the graph structure itself. If possible, I will also prove or disprove the conjecture that I present in Section 7 about the hyper Dirichlet process.

2 Background

Dawid and Lauritzen (1993) provide a general method for creating hyper Markov laws. They restrict their attention to decomposable graphs, which are particularly tractable (Frydenberg and Lauritzen, 1989). Decomposable graph can easily be 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 these distributions together by taking the base measure of one clique as a marginal, and conditioning the second clique on the intersection. They repeat this process for each clique, until all the cliques have been combined. The end result is a Markov distribution for the entire graph whose marginals are the provided base distributions over the cliques.

In a Bayesian setting, the distribution itself is random and therefore has a its own distribution called the prior law. Rather than specify base measures for each clique, one would specify marginal prior laws for the unknown distributions. Once again, it is required that these prior laws agree where the cliques intersect. These laws can be sewn together as in with the preceding paragraph to obtain a prior law for the entire graph.

As an example of the Dawid and Lauritzen (1993) construction, consider the problem of estimating the covariance matrix of a graphical Gaussian distribution. 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. 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, one arrives at the *hyper inverse Wishart*.

The hyper inverse Wishart distribution is one example of a hyper Markov distribution. It is a measure over $Q_{\mathcal{G}}$, the cone of real symmetric positive definite matrices, Σ, such that $\mathcal{N}(0, \Sigma)$ is Markov with respect to $\mathcal{G}$. The hyper Wishart distribution is a conjugate prior for a graphical Gaussian distribution 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. In contrast, non-parametric models make weak assumptions. Typical assumptions include continuity and the existence of some number of derivatives. Due to their weak assumptions, non-parametric priors have become popular in fields such as machine learning. My research aims to apply Markov and hyper Markov constraints to non-parametric models. This makes it possible to study graph selection problems without restricting attention to some relatively small parametric family. I begin with the Dirichlet Process, a commonly used non-parametric prior law. I then describe how to build this family into a non-parametric hyper Markov prior.

In my current work, I apply the framework of Dawid and Lauritzen (1993) to non-parametric priors. Instead of the inverse Wishart, the Dirichlet process prior is the base measure for each clique. Following the analogy, I build the marginals into a hyper Markov prior and call it the hyper Dirichlet process. Heinz (submitted) details the construction of this process and finds sufficient conditions to guarantee that it is hyper Markov. 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 modeling (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 with independence constraints.

Escobar and West (1995) use Dirichlet processes to estimate mixture models. The major benefit of this technique is that the number of mixing components does not need to be specified *a priori*. As an application of my theory, I extend this idea to Markov mixtures. I use a hyper Dirichlet process to create a mixture of hyper Markov distributions.

In the next section, I explain notation and review a few key concepts from graph theory. Following that, I discuss previous areas of research that form the motivation and foundation of my own work. In Section 4, I concentrate on the notions of Markov distributions and hyper Markov priors. In Section 5, I discuss the Dirichlet process, which has been studied extensively. In Section 6, I review one application of Dirichlet processes, the Dirichlet mixture of Gaussians. Following these sections I present my own research and future goals. In Section 7, I present the theory which I have already developed to generalize the Dirichlet process to the *hyper Dirichlet process*. As an application of this theory, I define the hyper Dirichlet mixture model in Section 8. I discuss my preliminary results from this model in Section 9. Finally, in Section 10, I present research questions which will guide my work. These are divided into two main categories: extending the current application, and studying hyper Markov generalizations for other non-parametric processes.

3 Definitions and Notation

Throughout this paper we consider a graph, $\mathcal{G}$, with vertex set (or node set) $\mathbf{V}$ and edge set $\mathbf{E}$. There is an edge from one vertex, γ_1, to another vertex, γ_2, if $(\gamma_1, \gamma_2) \in \mathbf{E}$. 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 proper 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 graph is *decomposable* if it 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 $\mathbf{S}_k = \mathbf{C}_k \cap \mathbf{H}_{k-1}$ and $\mathbf{R}_k = \mathbf{C}_k \setminus \mathbf{H}_{k-1}$. 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.

In graphical models, 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.

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

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.

4 Markov and Hyper Markov Measures

An undirected graph depicts the conditional independence structure for some variable X. Distributions that satisfy these constraints are called Markov measures.

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 for any decomposition $(\mathbf{A}, \mathbf{B})$,*

$$X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} | X_{\mathbf{A}\cap\mathbf{B}}[\theta],$$

where $X \sim \theta$.

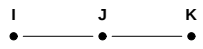

Figure 1: A graph depicting conditional independence of I and K given J.

For example, 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 $X \sim \theta$. It will be useful to keep Figure 1 in mind throughout this paper. While the graph visually has only three variables, it is representative of any connected graph of two cliques. Instead of one variable, let I, J, and K 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.

We denote the set of all distributions that are Markov on $\mathcal{G}$ by $\mathscr{M}(\mathcal{G})$. Dawid and Lauritzen (1993) showed that a probability measure is Markov if and only if it satisfies the *global Markov property*:

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

A Markov measure is determined by its clique marginals. If Q is a measure over $\mathcal{X}_{\mathbf{A}}$ and R is a measure over $\mathcal{X}_{\mathbf{B}}$, we say that they are *consistent* if they induce the same marginal over $\mathcal{X}_{\mathbf{A}\cap\mathbf{B}}$. In other words, if Q and R are consistent, then $Q_{\mathbf{A}\cap\mathbf{B}} = R_{\mathbf{A}\cap\mathbf{B}}$. Dawid and Lauritzen (1993) showed that under this condition, there is a unique distribution P, such that (i) $P_{\mathbf{A}} = Q$, (ii) $P_{\mathbf{B}} = R$, and (iii) $X_{\mathbf{A}} \perp\!\!\!\perp X_{\mathbf{B}} | X_{\mathbf{A}\cap\mathbf{B}}[P]$. We call P the *Markov combination* of Q and R, and denote it $P = Q \star R$.

In Bayesian statistics, we consider random measures. Thus, the measure has its own distribution called the *prior*. In this paper, we reserve the term *law* to refer to a distribution over measures. This is merely for clarity to simplify some definitions. For the prior law of a random measure, there exists a property that is similar to Definition 2.

Definition 3 HYPER MARKOV LAW. *If $\mathcal{L}$ is a law on $\mathscr{M}(\mathcal{G})$, we say it is* hyper Markov *on a decomposable graph, $\mathcal{G}$, if for any decomposition, $(\mathbf{A}, \mathbf{B})$,*

$$\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{A}\cap\mathbf{B}} \ [\mathcal{L}],$$

where $\theta \sim \mathcal{L}$.

We note that if the prior for θ is a hyper Markov law, then by definition, θ must be Markov. As with Markov laws, the hyper Markov property can restated as the *global hyper Markov property*: if $\mathbf{S}$ separates $\mathbf{A}$ and $\mathbf{B}$, then $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{S}}$. Note that a hyper Markov distribution may contain additional independence constraints not specified by the graph.

As with Markov measures, hyper Markov laws are determined by their clique marginals. Suppose each clique is endowed with a prior law for some random distribution. We say that the laws are *hyperconsistent* if they agree where the cliques intersect. Given two hyperconsistent laws, there is a natural way to combine them into a joint prior as shown in the next definition.

Definition 4 *Let Q over $\mathscr{M}(\mathcal{G}_{\mathbf{A}})$ and R over $\mathscr{M}(\mathcal{G}_{\mathbf{B}})$ be hyperconsistent laws. The* hyper Markov combination *of Q and R, denoted $Q \odot R$, is the unique law, $\mathcal{L}$, such that:*

1. $\mathcal{L}$ *is concentrated on* $\mathscr{M}(\mathcal{G}_{\mathbf{A}\cup\mathbf{B}})$,

2. $\mathcal{L}_{\mathbf{A}} = Q$,

3. $\mathcal{L}_{\mathbf{B}} = R$,

4. $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{A}\cap\mathbf{B}} [\mathcal{L}]$.

This law exists and is unique by Lemma 3.3 of Dawid and Lauritzen (1993). In the same paper, they generalize this to multiple cliques by iteratively forming hyper Markov combinations.

Suppose that the graph, $\mathcal{G}$, has a perfect ordering of cliques, $(\mathbf{C}_1, \ldots, \mathbf{C}_k)$. Further suppose that each clique is endowed with a marginal law, $\mathcal{M}_{\mathbf{C_i}}$, and that these laws are pairwise hyperconsistent. Define the following:

$$\mathcal{L}_{\mathbf{H}_1} = \mathcal{M}_{\mathbf{C}_1}, \tag{2}$$

$$\mathcal{L}_{\mathbf{H}_i} = \mathcal{L}_{\mathbf{H}_{i-1}} \odot \mathcal{M}_{\mathbf{C}_i}, \qquad \text{for } 1 < i \leq k. \tag{3}$$

Theorem 3.9 of Dawid and Lauritzen (1993) proves that $\mathcal{L} = \mathcal{L}_{\mathbf{H}_k}$ is the unique hyper Markov law whose clique marginals are the given hyperconsistent laws $(\mathcal{M}_{\mathbf{C}_i})$.

5 The Dirichlet Process

The Dirichlet process is a prior law, which provides a distribution over the space of probability distributions on $(\mathcal{X}, \mathcal{F})$. The Dirichlet process is non-parametric, meaning 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.

Definition 5 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*

$$(\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots, \mathbb{P}(A_k)) \sim \mathrm{Dir}(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k)), \tag{4}$$

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

5.1 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

$$\theta(\cdot) = \sum_{k=1}^{N} m_k \delta_{Z_k}(\cdot), \tag{5}$$

where the Z_k are independently distributed atoms from some distribution, G, and $\sum_{k=1}^{N} m_k = 1$ almost surely. The number of atoms, N, may be finite or infinite. The masses (or weights) are determined by successively breaking random pieces of a unit-length stick. Thus, $m_1 = t_1$, $m_2 = (1 - t_1)t_2$, and $m_k = t_k \prod_{i=1}^{k-1}(1 - t_i)$. For finite N, m_N is defined to be $1 - \sum_{i=1}^{N-1} m_i$, or equivalently $\prod_{i=1}(1 - t_i)$. If θ is a measure over $\mathcal{X}$, and $\mathbf{A} \subseteq \mathbf{V}$, then the marginal measure over $\mathcal{X}_{\mathbf{A}}$ is $\theta_{\mathbf{A}}$. For a stick-breaking measure, we write

$$\theta_{\mathbf{A}}(\cdot) = \sum_{k=1}^{N} m_k \delta_{Z_{k\mathbf{A}}}(\cdot), \tag{6}$$

where $Z_{k\mathbf{A}}$ is the marginal value of Z_k on $\mathbf{A}$.

Traditionally, stick-breaking measures are defined such that t_k is a Beta(a_k, b_k) random variable for $1 \leq k < N$. Thus, a stick-breaking measure is specified by the probability distribution, G, the number of atoms, N, and the countable sequence of Beta parameters $(a_k, b_k)_{k=1}^{N-1}$. Sethuraman (1994) showed that a Dirichlet Process, DP_α, 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 (stick-breaking representation). *Let* $\mathbf{A}$ *be any subset of* $\mathbf{V}$. *Let* G *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 measure (or distribution)* G *and precision* ν, *and write* $\theta \sim DP(\nu G)$, *if*

$$(\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots, \mathbb{P}(A_k)) \sim \text{Dir}(\nu G(A_1), \nu G(A_2), \ldots, \nu G(A_k)), \tag{7}$$

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

This definition is equivalent to Definition 5 by letting $\alpha = \nu G$. Here, ν and G are easily translated as the parameters of a stick-breaking measure. That is, the random atoms are iid G, 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 I prove, Definition 6 will be the definition of choice for much of this paper. The next theorem is an important result about Dirichlet processes.

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

(i) $X_i \sim G \ \ \forall\ i$.

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

For a proof, see Theorem 1.9.4 of Schervish (1995), p. 54.

The first property states that if the random measure is integrated out, then the marginal distribution of the data is $\overline{\alpha}$. This property implies that a Markov base measure ensures that the Dirichlet process, integrated over all possible θ, is a Markov distribution. However, this does not guarantee that $DP(\nu G)$ is a hyper Markov law. That requires the stronger condition that $\theta \sim DP(\nu G)$ is a Markov distribution with probability one. The second property states that if the prior law of θ is a Dirichlet process, then the posterior law is also a Dirichlet process, with an easily updated base measure. In Section 7, I use this property to show that a hyper Dirichlet process prior results in a hyper Dirichlet process posterior.

6 A Dirichlet Mixture of Gaussians

A parametric mixture model is of the form $f(\cdot) = \sum_{i=1}^k p_i f(\cdot|\pi_i)$, where $\sum_{i=1}^k p_i = 1$ and $\{f(\cdot|\pi)\}$ is some family of distributions indexed by a parameter π. Estimating a mixture model requires making inferences about the components (i.e. the π_i's) as well as the mixing weights of each (i.e. p_i). In some cases, the number of components, k, is unknown and must also be estimated. Dirichlet processes have been used in this area to handle all three problems simultaneously.

Escobar and West (1995) fit a Dirichlet mixture of Gaussians. In their model, the data $X_1, \ldots, X_n$ are conditionally independent and normally distributed, $X_i|\pi_i \sim N(\mu_i, V_i)$. The parameters are drawn from some prior distribution on $\mathbb{R} \times \mathbb{R}^+$. Having observed data, $D_n = \{x_1, \ldots, x_n\}$, the predictive distribution of Y_{n+1} is a Gaussian mixture specified by the posterior distribution of $\pi_{n+1}|D_n$. Calculation of this posterior is an example of Bayesian density estimation. When the shape of the prior is unknown, a Dirichlet process prior may be used, which results in a Dirichlet mixture of (Gaussian) distributions. We denote the prior law by $\mathcal{L} = DP(\nu G_0)$.

Denote by $\pi^{(i)} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$ the parameter values excluding π_i. The conditional prior law for π_i given the other parameters is

$$\pi_i|\pi^{(i)} \sim \frac{\nu}{\nu+n-1}G_0(\pi_i) + \sum_{j=1,j\neq i}^{n} \frac{1}{\nu+n-1}\delta_{\pi_j}(\pi_i). \tag{8}$$

It is clear from Equation 8 that there is a positive probability that π_i is equal to some other π_j $(i \neq j.)$ Therefore, the number of distinct values in π may be less than n. Denote the distinct values of π by $\pi^* = (\pi_1^*, \ldots, \pi_k^*)$, where the element π_j^* appears n_j times. The conditional prior can be rewritten as a mixture of $k+1$ components.

$$F(\pi_i|\pi^{(i)}) = \frac{\nu}{\nu+n}G_0(\pi_i) + \sum_{j=1}^{k} \frac{n_j}{\nu+n}\delta_{\pi_j^*}(\pi_i). \tag{9}$$

An immediate result of Equation 9 is that the conditional distribution of X_i given $\pi^{(i)}$ contains k Gaussian components and one non-Gaussian component, derived from the base measure of the Dirichlet process. Inference about the number of components is handled implicitly in the following sense. The parameters are unknown, but their posterior distribution is inferred from the data. In turn, this distribution implies a distribution over k, the number of components. Thus, the Dirichlet mixture yields a finite, but unknown number of mixture components. It is not necessary to specify k *a priori*.

Recall that the observations are conditionally independent given their parameters. As such, the predictive distribution of the next observation conditioned on the current sample and parameters satisfies $F(X_{n+1}|\pi, D_n) = \int F(X_{n+1}|\pi_{n+1})dF(\pi_{n+1}|\pi) = F(X_{n+1}|\pi)$. Bayesian density estimation is solved by integrating out the unknown π. This integral is intractable, but it can be estimated using a Gibbs sampler. The required conditional prior is given by

$$f(\pi_i|\pi^{(i)}, D_n) \propto \nu f_{G_0}(x_i)dG_i(\pi_i) + \sum_{j=1,j\neq i}^{n} f(x_i|\pi_j), \tag{10}$$

where f_{G_0} is the marginal distribution of x_i when $\pi_i \sim G_0$, and G_i is the posterior distribution of π_i given x_i when $\pi_i \sim G_0$ and $x_i|\pi_i \sim f(\cdot|\pi_i)$. Thus, the conditional prior is a mixture distribution:

$$F(\pi_i|\pi^{(i)}, D_n) = w_0 G_i(\pi_i) + \sum_{j=1,j\neq i}^{n} w_j \delta_{\pi_j}(\pi_i), \tag{11}$$

where the weights are $w_i = q_i / \sum_{j=0}^{k} q_j$, with $q_0 = \nu f_{G_0}(x_i)$ and $q_j = n_j f(x_i|\pi_j)$ for $j > 0$.

Equation 11 reveals the necessary conditions for creating a Gibbs sampler. First, the base measure must lead to a tractable calculation of $f_{G_0}(x_i) = \int f(x_i|\pi_i)dG_0(\pi_i)$. Second, we must be able to sample from the posterior update G_i. These conditions are met by using a conjugate prior as the base measure.

In the univariate model employed by Escobar and West (1995), G_0 is specified as a Normal $\times$ Inverse-Gamma distribution. The variance parameter V_i is Inverse-Gamma, where $V_i^{-1} \sim G(s/2, S/2)$, a gamma distribution with shape $s/2$ and scale $S/2$. Conditional on the variance, the mean has distribution $N(m, \tau V_i)$. This $IG \times N$ 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 parameter m, and scale parameter $M^{1/2}$, where $M = (1+\tau)S/s$.

7 The Hyper Dirichlet Process

Drawing from both the hyper Markov literature and previous results on Dirichlet processes, I construct a hyper Dirichlet process by considering a separate Dirichlet process for each clique. I begin with a graph consisting of two cliques, $\mathbf{A}$ and $\mathbf{B}$, and call the separator $\mathbf{C} = \mathbf{A} \cap \mathbf{B}$. I place a Dirichlet process prior over $\mathcal{X}_\mathbf{A}$, say $\mathcal{Q} = DP(\nu_1 Q)$, and another prior over $\mathcal{X}_\mathbf{B}$, say $\mathcal{R} = DP(\nu_2 R)$. The hyper Dirichlet process is a Markov measure such that the clique marginal measures are the specified Dirichlet processes. I summarize this construction below; the details of which are in a separate paper I've written (Heinz, submitted). I begin

by generalizing the definition of a Markov combination of distributions to general finite measures. Of course, I must require some type of consistency condition.

The marginal laws over $\mathcal{X}_{\mathbf{C}}$ are $\mathcal{Q}_{\mathbf{C}} = DP(\nu_1 Q_{\mathbf{C}})$ and $\mathcal{R}_{\mathbf{C}} = DP(\nu_2 R_{\mathbf{C}})$. To find the hyper Markov combination, I must ensure that these marginals are equivalent. From the stick-breaking construction, it is evident that this requires that $\nu_1 = \nu_2$ and $R_{\mathbf{C}} = Q_{\mathbf{C}}$. From this analysis, I define consistency of finite measures. The measures α over $\mathcal{X}_{\mathbf{A}}$ and β over $\mathcal{X}_{\mathbf{B}}$ are *consistent* if $\alpha(\mathcal{X}_{\mathbf{A}}) = \beta(\mathcal{X}_{\mathbf{B}})$ and $\overline{\alpha}_{\mathbf{C}} = \overline{\beta}_{\mathbf{C}}$. From here, I generalize the definition of a Markov combination.

Definition 8 Markov Combination of Finite Measures. *Suppose α and β are consistent finite measures. The Markov combination, denoted $\mu = \alpha \star \beta$, is*

$$\mu = \alpha(\mathcal{X}_{\mathbf{A}}) \cdot \left(\overline{\alpha} \star \overline{\beta}\right).$$

In the preceding paragraphs, I have shown that two DP measures are hyperconsistent if the base measures are consistent. In the stick-breaking construction, this means that the precisions are equal and the base distributions agree over $\mathbf{C}$. I now suppose that $\mathcal{Q}$ and $\mathcal{R}$ are hyperconsistent. By Lemma 3.3 of Dawid and Lauritzen (1993), this ensures the existence of a unique hyper Markov combination, $\mathcal{L} = \mathcal{Q} \odot \mathcal{R}$. It does not guarantee that $\mathcal{L}$ is a Dirichlet process. Since the Dirichlet process is so well-studied, it is useful to determine conditions that ensure this property.

Let $\mathcal{D} = DP(\nu G)$ and note that $\mathcal{D}_{\mathbf{A}} = \mathcal{Q}$ and $\mathcal{D}_{\mathbf{B}} = \mathcal{R}$. Therefore, if it is hyper Markov, it must be that $\mathcal{D} = \mathcal{Q} \odot \mathcal{R}$ by Definition 4. The problem now reduces to determining when this process is hyper Markov. For insight, I consider a case in which it is not hyper Markov.

The hyper Markov property requires in part that if $\theta \sim \mathcal{L}$, then $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$ almost surely $[\mathcal{L}]$. Consider the case in which $Q_{\mathbf{A}\setminus\mathbf{B}}$ and $R_{\mathbf{B}\setminus\mathbf{A}}$ are continuous measures, but $Q_{\mathbf{C}} = R_{\mathbf{C}}$ is degenerate at some point c. Clearly, $\theta_{\mathbf{C}} = \delta_c$ with probability one. Recall from Equation 6 that $\theta_{\mathbf{A}} = \sum_{k=1}^{\infty} m_k Z_{k\mathbf{A}}$ and $\theta_{\mathbf{B}} = \sum_{k=1}^{\infty} m_k Z_{k\mathbf{B}}$, for some random $\vec{m}$ whose elements sum to 1 and random atoms Z_k, where $Z_{k\mathbf{A}}$ and $Z_{k\mathbf{B}}$ are marginal values of Z_k. Since $R_{\mathbf{B}}$ is continuous, each $Z_{k\mathbf{B}}$ is distinct almost surely. Therefore, if we observe $\theta_{\mathbf{B}}$, then we know $\vec{w}$ modulo permutations. Obviously, these weights provide information about the value of $\theta_{\mathbf{A}}$. Indeed, we now know every mass in the pmf of $\theta_{\mathbf{A}}$, though we do not know the location of these masses. Furthermore, this information is not known if we only observe $\theta_{\mathbf{C}}$, whose degenerate distribution gives no information about either $\theta_{\mathbf{A}}$ or $\theta_{\mathbf{B}}$.

As this counterexample shows, I want to ensure that $\theta_{\mathbf{B}}$ does not provide information about the weights that is not given by observing $\theta_{\mathbf{C}}$ alone. I have discovered that I can do so by imposing a constraint on R.

Refinement Condition:

$$Z_{k\mathbf{C}} = Z_{k\mathbf{C}} \Longrightarrow Z_{k\mathbf{B}} = Z_{k\mathbf{B}} \quad \forall k,\ a.s.[R]. \tag{12}$$

The basic idea behind this theory is that whenever two of the random atoms coincide, information about the weights is lost. The corresponding weights are no longer observed. Instead, we observe only their sum. The Refinement Condition ensures that if the information is lost from $\theta_{\mathbf{C}}$, then it is also lost from $\theta_{\mathbf{B}}$. In my other paper, I show that this condition is sufficient to imply that $DP(\alpha Q \star R)$ is hyper Markov as long as $Q \star R$ is well-defined.

This condition is sufficient, but it is clearly not necessary. To see this, note that conditional independence holds by symmetry if we replace $\mathbf{B}$ with $\mathbf{A}$ and R with Q. For example, suppose $Q_{\mathbf{A}} = \delta_a$ for some a. For $\theta \sim \mathrm{DP}(\alpha Q \star R)$, the marginal $\theta_{\mathbf{A}}$ is almost surely δ_a. Thus, it is trivially true that $\theta_{\mathbf{A}} \perp\!\!\!\perp \theta_{\mathbf{B}} | \theta_{\mathbf{C}}$, even if the Refinement Condition (as written) does not hold. In this counterexample, the refinement condition holds for $Q_{\mathbf{A}}$ and $Q_{\mathbf{C}}$ rather than $R_{\mathbf{A}}$ and $R_{\mathbf{C}}$. I conjecture that the necessary and sufficient condition for hyper Markovity is that either Q or R satisfies the Refinement Condition, but I do not know the proof. A side goal of my research plan is to prove or disprove this conjecture if possible.

At first glance, the Refinement Condition may appear unduly restrictive, but it is actually very general. As an important example, if the base measures are continuous, then it holds trivially. When the Refinement Condition *is* satisfied, I call the hyper Markov combination a *hyper Dirichlet process* and write $\theta \sim HDP(\nu G)$ to emphasize this. A key aspect of this construction is that the hyper Dirichlet process is also a standard Dirichlet process. Hence, my research has uncovered a non-parametric process that takes advantage of a

conditional independence structure and there is already a wealth of information about it. In particular, many applications which rely on Dirichlet processes can be generalized to use a hyper Dirichlet process in graphical cases. For example, I can create a mixture model of graphical components as in the next section.

Another useful property of the hyper Dirichlet process prior is that hyper Markovity will persist in the posterior if the Refinement Condition is satisfied. Consider $(X_1, \ldots, X_n)$, a sample of n independent observations of θ, where $\theta \sim HDP(\nu G)$. From Theorem 7, we see that the posterior is $DP(\nu' G')$, where $G' = \alpha G + \sum_{i=1}^{n} \delta_{X_i}$ and $\nu' = \nu + n$. Because the Refinement Condition holds for G, the new points of mass, (δ_{X_i}) will be constrained so that G' also satisfies the Refinement Condition. The details of this proof are omitted for space, but they can be found in Heinz (submitted).

8 Hyper Dirichlet Mixtures

My work extends the Escobar and West (1995) model to allow multivariate observations as well as the specification of a graphical model. In this way, my model is at the confluence of the hyper Markov prior literature and the Bayesian density estimation literature. It is a hyper Dirichlet mixture of graphical Gaussian distributions. In other words, my model is a mixture of components, each of which is a Markov distribution.

I begin by assuming that the graph, $\mathcal{G}$, is known and has cliques $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ and separators $\mathcal{S} = \{\mathbf{S}_2, \ldots, \mathbf{S}_k\}$. The set of all Markov distributions for $\mathcal{G}$ is denoted by $\mathscr{M}(\mathcal{G})$. The family of p-variate Normal distributions is $\mathscr{N}_p = \{N(\mu_p, \Sigma)\}$, where p is the number of nodes on $\mathcal{G}$. The graphical Gaussian model, $\mathscr{N}_\mathcal{G} = \mathscr{N}_p \cap \mathscr{M}(\mathcal{G})$, is the sub-family of Normal distributions which are Markov with respect to $\mathcal{G}$. A prior distribution for this family is a measure over $\mathbb{R}^p \times Q_\mathcal{G}$, where $Q_\mathcal{G} = \{\Sigma : N(0, \Sigma) \in \mathscr{N}_\mathcal{G}\}$. I specify a hyper Dirichlet mixture by

$$\begin{aligned} (\mu_i, V_i) &\sim HDP_\mathcal{G}(\nu G_0) \\ (X_i | \mu_i, V_i) &\sim N_p(\mu_i, V_i), \end{aligned}$$

where G_0 is a distribution on $\mathbb{R}^p \times Q_\mathcal{G}$. I generalize Escobar and West's $IG \times N$ prior to a hyper inverse Wishart $\times$ hyper Normal ($HIW \times HN$) prior. That is, if $(\mu, V) \sim G_0$, then

$$V \sim HIW_\mathcal{G}(d, D) \tag{13}$$

$$(\mu | V) \sim N_p(m, \tau V) \tag{14}$$

$$(X | \mu, V) = N_p(\mu, V), \tag{15}$$

where $HIW_\mathcal{G}$ is the hyper Inverse Wishart, hyper Markov with respect to $\mathcal{G}$, having d degrees of freedom and location D. I write the density of the hyper inverse Wishart in terms of the clique and separator marginal densities. For a $p \times p$ matrix, V, let $V_{\mathbf{AB}}$ denote the sub-matrix of V with rows in $\mathbf{A}$ and columns in $\mathbf{B}$. The density of the $HIW(d, D)$ distribution is

$$dHIW(V; d, D) = \frac{\prod_{j=1}^{k} dIW(V_{\mathbf{C_j C_j}}; d, D_{\mathbf{C_j C_j}})}{\prod_{j=2}^{k} dIW(V_{\mathbf{S_j S_j}}; d, D_{\mathbf{S_j S_j}})}, \tag{16}$$

where $dHIW(x; d, D)$ is the density of the saturated inverse Wishart distribution evaluated at x with d degrees of freedom and location D.

I will now show that G_0 is a hyper Markov prior, which implies that the model specified by (13) - (15) is a mixture of Markov distributions.

The hyper Inverse Wishart is a distribution for $V \in Q_\mathcal{G}$. It is a hyper Markov prior for the graphical Gaussian with known mean (Dawid and Lauritzen, 1993; Letac and Massam, 2007). By definition, this ensures that $N_p(m, \tau V)$ is a Markov distribution for μ_i. As such, we can express the density in terms of the marginal densities of the cliques and separators, as in Equation 16.

$$dHN(\mu; m, V) = \frac{\prod_{j=1}^{k} dN(\mu_\mathbf{C}; m_{\mathbf{C_j}} V_{\mathbf{C_j C_j}})}{\prod_{j=2}^{k} dN(\mu_{\mathbf{S_j}}; m_{\mathbf{S_j}}, V_{\mathbf{S_j S_j}})}, \tag{17}$$

where $dN(x; m, V)$ is the Normal density evaluated at x with mean m and variance V.

Because $V \in Q_{\mathcal{G}}$, G_0 is a prior law over $\mathscr{M}(\mathcal{G})$. Let $(\mathbf{A}, \mathbf{B})$ be a decomposition of $\mathcal{G}$ with $\mathbf{S} = \mathbf{A} \cap \mathbf{B}$. The density of G_0 is

$$\begin{aligned} & g(\mu, V) \\ &= g(V)g(\mu|V) \\ &= \frac{dHIW(V_{\mathbf{AA}}; d, D_{\mathbf{AA}})dHIW(V_{\mathbf{BB}}; d, D_{\mathbf{BB}})}{dIW(V_{\mathbf{SS}}; d, D_{\mathbf{SS}})} \cdot \frac{dHN(\mu_{\mathbf{A}}; m_{\mathbf{A}}, V_{\mathbf{AA}})dHN(\mu_{\mathbf{B}}; m_{\mathbf{B}}, V_{\mathbf{BB}})}{dN(\mu_{\mathbf{S}}; m_{\mathbf{S}}, V_{\mathbf{SS}})} \end{aligned} \tag{18}$$

Noting that $\mathbf{S} \subseteq \mathbf{B}$, we also have

$$f(\mu_{\mathbf{B}}, V_{\mathbf{BB}}, \mu_{\mathbf{S}}, V_{\mathbf{SS}}) = f(V_{\mathbf{BB}})f(\mu_{\mathbf{B}}|V_{\mathbf{BB}}). \tag{19}$$

Equations 18 and 19 together imply

$$\begin{aligned} f(\mu_{\mathbf{A}}, V_{\mathbf{AA}}|\mu_{\mathbf{B}}, V_{\mathbf{BB}}, \mu_{\mathbf{S}}, V_{\mathbf{SS}}) &= \frac{f(\mu, V)}{f(\mu_{\mathbf{B}}, V_{\mathbf{B}})} \\ &= \frac{dHIW(V_{\mathbf{AA}}; d, D_{\mathbf{AA}})dHN(\mu_{\mathbf{A}}; m_{\mathbf{A}}, V_{\mathbf{AA}})}{dIW(V_{\mathbf{SS}}; d, D_{\mathbf{SS}})dN(\mu_{\mathbf{S}}; m_{\mathbf{S}}, V_{\mathbf{SS}})} \\ &= f(\mu_{\mathbf{A}}, V_{\mathbf{AA}}|\mu_{\mathbf{S}}, V_{\mathbf{SS}}), \end{aligned}$$

which implies that G_0 is a hyper Markov prior.

In addition to being hyper Markov, G_0 is also conjugate. Consider observing a random variable, X, where $X \sim N(\mu, V)$ and the parameters $(\mu, V) \sim G_0$ 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 Bayesian 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.

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

Recall that the density of the hyper inverse Wishart can be expressed in terms of clique and separator marginals, as in Equation 16. Likewise, the posterior can be found by updating each clique and separator individually. Thus, the posterior distribution of V after observing D_n is

$$(V|X) \sim \frac{\prod_{c \in \mathcal{C}} dIW(V_c; d + n, D_c + \Phi_c)}{\prod_{s \in \mathcal{S}} dIW(V_s; d + n, D_s + \Phi_s)}, \tag{20}$$

where $\Phi = \sum_{i=1}^{n} x_i x_i'$ is n times the sample covariance matrix. Therefore, the posterior distribution of $(V|X)$ is $HIW(d + n, D + \Phi)$.

Taking further advantage of the hyper Markov structure, I 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

$$X \sim \frac{\prod_{c \in \mathcal{C}} dT\left(X_c; d + 1 - |c|, m_c, \frac{\tau+1}{d+1-|c|} D_c\right)}{\prod_{s \in \mathcal{S}} dT\left(X_s; d + 1 - |s|, m_s, \frac{\tau+1}{d+1-|s|} D_s\right)}, \tag{21}$$

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. I 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 21 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 my 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, I need to consider cases in which $m \neq 0$. This is what led me to discover the more general form of the hyper t.

Recall that there are two requirements for solving the Bayesian density estimation problem using a Gibbs sampler. First, the marginal density of X must be tractable. Secondly, we must be able to generate the posterior distribution of $(\mu, V|X)$. My model meets both of these requirements. The marginal density is easily calculated using Equation 21. The most difficult step of the algorithm is to generate a hyper inverse Wishart random variable. However, this can be done following the method of Carvalho et al. (2007). Essentially, the algorithm begins by generating an inverse Wishart for $V_{\mathbf{C}_1}$. For $i > 1$, $V_{\mathbf{C}_i}$ is generated as a conditional inverse Wishart, given that the elements of $V_{\mathbf{S}_i}$ are known. Since the clique sub-matrices are sufficient for V, the other values can be computed analytically. The posterior density estimate can be sampled using the following Gibbs algorithm.

1. Choose initial values for each π_i. A suitable method is to draw from the posterior distribution given X_i.
2. For $i = 1, \ldots, n$: Sample (μ_i, V_i) according to Equation 11, using the current values of μ_j and V_j for $j \neq i$.
   (a) Set $q_0 = \alpha d HT(X_i; d+1, m, (\tau+1)D)$, and $q_j = n_j d N(X_i; \mu_j, V_j)$.
   (b) Reweight the q_j's to sum to one.
   (c) With probability q_j, set $(\mu_i, V_i) := (\mu_j, V_j)$. Otherwise draw a new value from G_0.
3. Repeat Step 2 until convergence.

9 Preliminary Results

To test my hyper Dirichlet mixture model, I simulated Gaussian data with the graph [12][23]. I specified the hyperparameters, which I used to generate component distributions. I then generated observations according to a preset weighting of the components. The data set consisted of 300 observations and 3 mixing components. Each observation had 3 variables. The simulation employed the following algorithm:

1. Specify the parameters for $HIW_{\mathcal{G}}(D, d)$.
2. Specify the center (m) and spread (τ) of the component means
3. Specify the number of components and their relative weights $(p_1 \ldots p_k)$.
4. For each component, $i = 1 \ldots k$:
   (a) Generate a covariance matrix, $V_i \sim HIW_{\mathcal{G}}(D, d)$.
   (b) Generate a mean, $\mu_i \sim N(0, \tau V_i)$.
5. For each observation, $j = 1 \ldots n$:
   (a) Randomly choose a component, i, according to the weights.
   (b) Generate the observation, $X_j \sim N(\mu_i, V_i)$.

A subtle difficulty in this algorithm is specifying D such that $HIW_{\mathcal{G}}(D, d)$ is really hyper Markov. I achieve this by specifying the clique submatrices, which are the sufficient statistics for the graphical Gaussian distribution. When the edge (a, b) is missing, then the covariance element D_{ab} can be analytically calculated from the other values. A straightforward way to calculate the full matrix is to invert the clique sub-matrices, combine them to find D^{-1}, then invert once more to find D (Roverato, 2000; Carvalho et al., 2007). Denote

the clique submatrices by $D_{C_iC_i}$ and the separator submatrices by $D_{S_iS_i}$. The inverse covariance matrix can be computed by

$$D^{-1} = \sum_{i=1}^{C} [D^{-1}_{C_iC_i}]_0 - \sum_{i=2}^{C} [D^{-1}_{S_iS_i}]_0, \tag{22}$$

where the notation $[A]_0$ means to extend the matrix with zeroes to the full size. A similar transformation reduces the computational burden when trying to invert a matrix in $Q_{\mathcal{G}}$. This is another way in which a graphical structure reduces the complexity of model estimation.

In my simulation, the location parameter for the hyper Wishart distribution was

$$D = \begin{pmatrix} 1 & .5 & \mathbf{.1} \\ .5 & 1 & .2 \\ \mathbf{.1} & .2 & 1 \end{pmatrix},$$

where the bold values are determined analytically from the clique sub-matrices. As a simple check, one can see that in the inverse-covariance matrix (D^{-1}), the bold elements are 0, corresponding to the absence of an edge in the graphical model.

I generated three graphical components according to this model. The means and covariance matrices are given in Table 1. Since the component for each data point was random, the actual mixing weights in the sample vary from the theoretical weights. This is shown in Table 2.

i	p_i	μ_i	V_i
1	$\frac{8}{15}$	$\begin{pmatrix} 0.116 \\ -0.349 \\ 0.402 \end{pmatrix}$	$\begin{pmatrix} 0.134 & 0.0505 & 0.00364 \\ 0.0505 & 0.0884 & 0.00637 \\ 0.00363 & 0.00637 & 0.0522 \end{pmatrix}$
2	$\frac{5}{15}$	$\begin{pmatrix} -0.123 \\ 0.205 \\ 0.290 \end{pmatrix}$	$\begin{pmatrix} 0.125 & 0.0149 & 0.00371 \\ 0.0149 & 0.0459 & 0.0114 \\ 0.00371 & 0.0114 & 0.0728 \end{pmatrix}$
3	$\frac{2}{15}$	$\begin{pmatrix} -0.00218 \\ -0.0119 \\ -0.496 \end{pmatrix}$	$\begin{pmatrix} 0.190 & 0.124 & 0.0344 \\ 0.124 & 0.158 & 0.0439 \\ 0.0344 & 0.0439 & 0.0869 \end{pmatrix}$

Table 1: Randomly generated Normal mixture components for the small simulation

i	n	Actual	Theoretical
1	164	.547	.533
2	110	.367	.333
3	26	.087	.133
Tot	300	1	1

Table 2: Actual and theoretical mixing weights for the small simulation

I initialized the Gibbs sampler by putting each point in its own component with parameters drawn randomly from the posterior, $f(\mu_i, V_i|X_i)$. Thus, the 0^{th} sample is a mixture of n Markov distributions, one for each data point. I chose a burn-in of 3000 iterations then counted the next 9000 iterations as the posterior sample. A major goal of my research plan is to find a method to determine the appropriate burn-in period.

The first statistic I examined was the number of estimated components at each Gibbs draw. The number of components ranged from 3 to 10. The median was 5, which accounted for 37% of the samples. Samples with 4 components and samples with 6 components accounted for about 25% each, for a total of 87% within 5 ± 1. This is a great reduction from 300 data points to about 5 components, however the actual data is

from a mixture of only 3 Normals. Thus, the initial estimate of the number of components is too high. I investigated this further by plotting the component means against the sequence number of the Gibbs draw. The plot is shown in Figure 2. The horizontal axis is the first dimension of the mean vector (μ_{i1}) for each component. The vertical axis is the sequence number starting after the burn-in. In other words, each row has a point for each component corresponding to the mean of X_1 under that component distribution. This plot takes advantage of the fact that μ_{i1} is almost surely unique for each component. Thus, if two samples show the same value of μ_{i1} then they are from the same component. By following the plot from bottom to top, one can see at a glance how the component means change as the Gibbs sampler evolves. In particular, a long vertical line shows that a component persisted through many iterations. One notable fact is that when I rank the components by this persistence, there is a large difference after the top three components before the fourth component. Thus for a long time, the sampler is choosing the same three components and trying different possible components, but discarding them right away. In a sense, there are three major components. This is ostensibly good because I should expect three components. Unfortunately, the persistence of these values through so many trials means one of two things. Either the sampler had not converged yet, or even after convergence the sampler has a strong autocorrelation. The latter problem means that I can only keep every c^{th} output to get an independent sequence. The former problem is remedied by using a longer burn-in period. This problem is less cumbersome since the solution does not increase with the number of useable draws needed.

The Gibbs sampler appears to put too much weight on the data and not enough on the prior. That is, there is a very high chance to keep components from iteration to iteration. The problem is not enough emphasis on the Dirichlet prior, so that previously seen values are too likely. This can be resolved by choosing good hyperparameters, which is one of my research goals. As I discuss in the next section, one idea is to add an additional tier to the hierarchical model to allow D, d, m, and τ to be determined via sampling.

Another aspect of the data that I was really curious about was the relative weights of each cluster. In particular, I was curious if the weights corresponded to the actual representativeness of each cluster in the simulated data. It is somewhat difficult to find a good way to visualize this for a variety of reasons. One problem is that the weight vector lies in some moderately high-dimensional simplex. Even considering this, there is an additional complexity since the number of dimensions changes from sample to sample. Finally, there is the matter of labeling, though this is trivial compared to the other issues. That is, a component that persists from one observation to the next may be component 1 in one draw, and component 3 in another. My solution is to sort the weights and look at side-by-side boxplots of the ranked component weights. This shows what the largest component weight is in each sample, down to the smallest weight. In practice, I only looked at the top 5 weights, since the remaining weights accounted for a very small proportion of the data. The boxplots are shown in Figure 5. The horizontal lines are the correct weights from the three simulated components. By looking at the medians, we see that the median of the largest component decently approximates the actual weight of the largest component. The second- and third-largest components sum to a reasonable approximation of the second actual weight. Although admittedly weak evidence, this indicates that perhaps the second component is being split into two separate components. Unfortunately, the individual data points in the simulation are not tagged to show which Normal component they come from. This would be a good way to see if the Gibbs sampler is doing a good job separating the sample points by their component distribution. The next simulation I run will have this improvement.

10 Further Ideas

10.1 Current Model

A major unresolved question about hyper Dirichlet mixtures is how many dimensions can be handled realistically. Preliminary runs of the Gibbs sampler have exhibited some difficulty in mixing components. A way to encourage mixing would be to use a prior for k, the number of unique parameter values. I can specify the prior implicitly by placing a distribution on α, the prior precision for the hyper Dirichlet process. Antoniak (1974) shows that the number of unique draws from a Dirichlet process depends on α and n, however, it is unclear how this extends to Dirichlet mixtures. In a Dirichlet process, if $n-1$ observations are known, then the probability of the n^{th} observation being unique is $\alpha/(\alpha+n)$. This implies a distribution

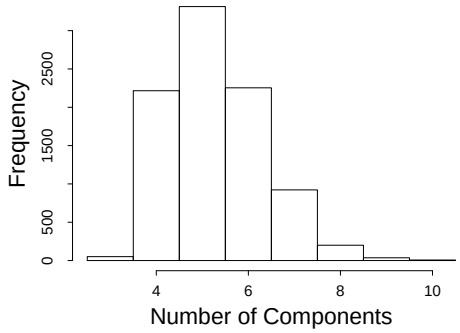

Figure 2: The frequency of k, the number of components, in a Gibbs sample of 9000 for the small simulation

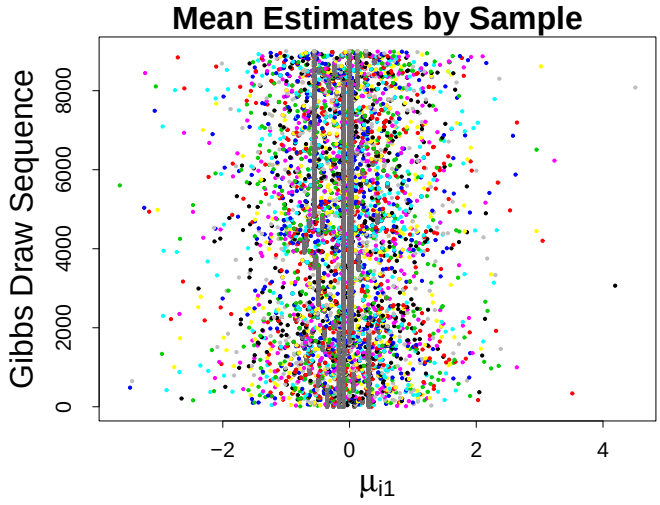

Figure 3: A plot of the components versus the sample number

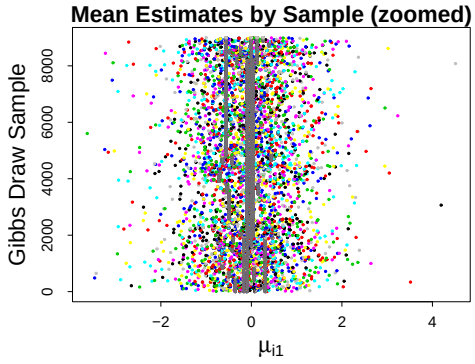

Figure 4: A plot of the components (zoomed)

for k. This differs from a Dirichlet process mixture in which x_i is also observed. The probability of a unique draw, w_0 in Equation 11, depends on H_0 and $f(X_i|\pi_j)$ in addition to α and n. For example, if X_i is close to X_j, then the probability that $\pi_i = \pi_j$ is increased. This shows the importance of dimensionality. The definition of "close" is not constant as the dimension increases. The relationship between the base measure, α, and the induced distribution on k will be interesting to investigate. This is related to my next research goal.

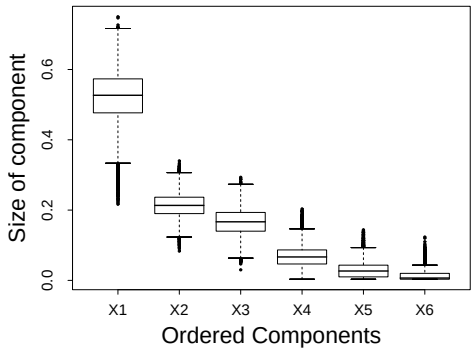

Figure 5: Side-by-Side Boxplots of sorted weight values

I would like to incorporate learning about the hyperparameters into the algorithm. Escobar and West (1995) extend their model by including a prior for m and τ. I can extend my model in the same way, using an inverse Gamma prior for τ and a conditional prior for $(m|\tau)$ from the Normal family. Furthermore, I want to find a suitable prior for d and D. In the extended model, α and d are the most interesting parameters because they most influence the distribution for the number of components. α influences the overall probability of repeating components before observing X_i. The degrees of freedom parameters controls the scale of the hyper t-distribution. For small d the distribution is more concentrated around m, which leads to a relatively large probability for drawing a new atom. This is assuming that m is near the majority of observations as it should be. The hyperparameter τ controls how widely spread the component means are. If τ is small relative to the spread of the data, then we expect the components to be similar.

When the hyper Dirichlet mixture is upgraded to estimate hyperparameters, it will be very useful to fit Gaussian mixtures given a graphical model. Beyond this, I will consider problems in which the independence structure is unknown. My first attempt will be to incorporate the kind of MCMC algorithm described by Giudici and Green (1999). They propose an algorithm which incorporates dropping and adding edges to the graph. I would like extend their work to my model, which would result in an algorithm that is able to determine the structure of a mixture of graphical Gaussian models in addition to fitting the components. In summary, the algorithm would incorporate ideas from three different research areas: hyper Markov models, covariance (graph) selection, and non-parametric Bayesian density estimation. The end product will be a non-parametric model with the only major assumption being that the data can be represented as a mixture of Gaussians. Even this assumption can be relaxed by considering other distributions. For example, if the data distribution exhibits heavy tails, a mixture of t-distributions could be estimated instead of Gaussians. Another possible generalization is to replace the HIW prior for V_i with the more general class of inverse Wisharts presented by Letac and Massam (2007).

Another goal is to develop good diagnostics to indicate that the Gibbs sampler converged and that it converged to the correct answer. The available methods include predicting new observartions and cross-validation. I also hope that my model shows self-consistency. I can demonstrate this by simulating data from the fit of a real data sample. If the Gibbs sampler is well-behaved, then the simulated data should look like the real data, and the model fit for simulated data should be close to the original fit.

10.2 Summary of Current and Proposed Research

To summarize, my work has already shown how hyper Markov priors can work in non-parametric settings. I have found sufficient conditions for a Dirichlet process to be hyper Dirichlet. The theory of hyper Dirichlet processes provides the benefits of the hyper Markov structure without requiring knowledge of the shape of the prior law. My construction also benefits from previous research on standard Dirichlet processes. The application I have studied to date is a mixture of Gaussian components, each of which is Markov for a given graph. This model can be fit using a Gibbs sampler which I have programmed and tested in R.

My proposed research goals are as follows:

- An immediate goal will be to translate the Gibbs sampler from R to C so that it can handle larger data sets and run more efficiently. This is necessary to explore my research goals in a timely manner.
- I will investigate ways to determine the necessary burn-in period for the Gibbs sampler to converge. A related question is to determine the autocorrelation of the observations after convergence. A large correlation means that the sampler should only accept every c^{th} iteration to obtain an iid sample from the posterior.
- I will expand the Gibbs sampler to include learning about hyperparameters. I will do this by adding a new tier to my hierarchical model with a prior for the hyperparameters, D, d, m and τ.
- I will examine the feasibility of learning about the graphical structure by adding or dropping edges as part of the Gibbs sampler.
- I have a large data set that was previously used to investigate covariance selection (Levina et al., 2008). I will use my sampler to estimate my model on this data set.
- I will investigate the tradeoff between my model and other estimation techniques, such as kernel smoothing. I hope to have an algorithm that is fast as well as accurate.
- I will develop good diagnostics to test fits of my model.
- If possible, I would like to determine the necessary conditions for this process to be a hyper Markov law. For a perfect ordering of cliques, I conjecture that for all $i > 2$, the Refinement Condition must hold for the i^{th} history or the i^{th} clique.

10.3 Other Processes, Other Problems

There are other problems that I could investigate using a hyper Dirichlet process. These are ideas which are out of the scope of my current work. Nevertheless, these ideas serve as examples for future work which stems from what I have already accomplished. For example, rather than estimating parameters for Gaussian distributions, I could estimate the "scale of membership" in k classes for n data points. This is a similar idea to block modeling and GoM models, as in Airoldi et al. (2008). In a block model, observations are assumed to be independent given the group membership for the data. This is akin to the mixture models alluded to above. Grade of Membership (GoM) models extend this to allow group membership to be divided for each observation. For example, X_i may have 20% membership in groups 2 and 4, and 60% membership in group 7. This leads to a mixture model in which the mixing weights are allowed to vary by observation. A third way to achieve mixing would be to consider what I call a "scale of membership". This is the GoM model which drops the constraint that the total membership grade for each point is 1. Thus, some points can be very representative of many classes, while others are not members of any groups. The latent variable of interest is $X_i = (X_{i1}, \ldots, X_{ik})$, where X_{ij} represents how well observation i is represented by class j. A hyper Dirichlet process could be used to model the distribution of $X_1, \ldots, X_n$. Given X_i, the observation could be modeled by an additive model such as linear regression.

The "scale of membership" idea has been used in machine learning literature with the constraint $X_{ij} \in \{0, 1\}$. Such processes are known as *Beta processes* or *Indian buffet process* (Thibaux and Jordan, 2007). In this setting, instead of groups or components, there are features. X_{ij} records whether or not observation i has feature j. The generalization to a *hyper Beta process* is to think about independence constraints on the different features. That is, the variable $X_i = \{X_{i1}, \ldots, X_{ik}\}$ is a latent variable from a k-way contingency table. Similar to the Dirichlet mixture, the beta process does not require us to choose the number of features *a priori*. This is a natural extension of my earlier theory because a Beta process can also be written as a countable sum of weighted random variables.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed-memberhsip stochastic blockmodels. *Journal of Machine Learning Research*, **9** 1823–1856.

Allcroft, D. J. and Glasbey, C. A. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Applied Statistics*, **52** 487–498.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, **2** 1152–1174.

Banerjee, O., Laurent and Ghaoui, E. (2007). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*. URL `http://www.princeton.edu/ aspremon/CovSelSIMAX.pdf`.

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. URL `http://ftp.stat.duke.edu/WorkingPapers/05-03.html`.

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.

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

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

Frydenberg, M. and Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, **76** 539–555.

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

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

Heinz, D. (submitted). Building hyper dirichlet processes for graphical models. *Electronic Journal of Statistics*.

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

Koul, H. and Deshpande, J. (eds.) (1995). *Analysis of Censored Data*.

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

Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, **2** 245–263.

Meila, M. and Shen, X. (eds.) (2007). *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*. URL `http://www.stat.umn.edu/ aistat/proceedings/start.htm`.

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.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, **87** 99–112.

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

Sebastiani, M. R. (2003). Markov random-field models for estimating local labour markets. *Applied Statistics*, **52** 201–211.

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.

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.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In Meila and Shen (2007). URL `http://www.stat.umn.edu/ aistat/proceedings/data/papers/071.pdf`.

Zachary, S. and Ziedins, I. (1999). Loss networks and Markov random fields. *Journal of Applied Probability*, **36** 403–414.