

A SIMPLE EXAMPLE OF DIRICHLET PROCESS MIXTURE INCONSISTENCY FOR THE NUMBER OF COMPONENTS

JEFFREY W. MILLER AND MATTHEW T. HARRISON

ABSTRACT. For data assumed to come from a finite mixture with an unknown number of components, it has become common to use Dirichlet process mixtures (DPMs) not only for density estimation, but also for inferences about the number of components. The typical approach is to use the posterior distribution on the number of components occurring so far — that is, the posterior on the number of clusters in the observed data. However, it turns out that this posterior is not consistent — it does not converge to the true number of components. In this note, we give an elementary demonstration of this inconsistency in what is perhaps the simplest possible setting: a DPM with normal components of unit variance, applied to data from a “mixture” with one standard normal component. Further, we find that this example exhibits severe inconsistency: instead of going to 1, the posterior probability that there is one cluster goes to 0.

1. INTRODUCTION

It is well-known that Dirichlet process mixtures (DPMs) of normals are consistent for the density — that is, given data from a sufficiently regular density p_0 the posterior converges to the point mass at p_0 (see [17, 4] for details and references). However, it is easy to see that this does not imply consistency for the number of components, since for example, a good estimate of the density might include superfluous components having vanishingly small weight.

Despite the fact that a DPM has infinitely many components with probability 1, it has become common to apply DPMs to data assumed to come from a finite mixture, and to apply the posterior on the number of components used to generate the observed data (in other words, the posterior on the number of clusters in the data) for inferences about the true number of components (see [1, 16, 14, 10, 8, 18, 9] for a few prominent examples). Thus, it is important to understand the properties of this procedure.

In this note, we give a simple example in which a DPM is applied to data from a finite mixture and the posterior distribution on the number of clusters does not converge to the true number of components. In fact, DPMs exhibit this type of inconsistency under very general conditions, as we will show elsewhere — however, the aim of this note is brevity and clarity. To this end, we focus our attention on a special case that is as simple as possible: a “standard normal DPM”, that is, a DPM using univariate normal components of unit variance, with a standard normal base measure (prior on component means).

Division of Applied Mathematics, Brown University, Providence, RI 02912.
jeffrey_miller@brown.edu.

Some authors have empirically observed that the DPM posterior tends to overestimate the number of components (e.g. [16, 9, 13], among others), and have found that ignoring tiny clusters tends to mitigate this issue. It might be possible to obtain consistent estimators in this way. However, by adopting such a procedure, one is abandoning the DPM model, and it is not clear what model (if any) would give rise to such a procedure.

A more natural way to obtain consistency is the following: if the number of components is unknown, put a prior on the number of components. For example, draw the number of components s from a probability mass function $p(s)$ on $\{1, 2, \dots\}$ with $p(s) > 0$ for all s , draw mixing proportions $\pi = (\pi_1, \dots, \pi_s)$ from an s -dimensional Dirichlet (given s), draw component parameters $\theta_1, \dots, \theta_s$ i.i.d. (given s and π) from an appropriate prior, and draw X_1, X_2, \dots i.i.d. (given s , π , and $\theta_{1:s}$) from the resulting mixture. This approach has been widely used [11, 15, 6, 12]. Strictly speaking, as defined, such a model is not identifiable — but it is fairly straightforward to modify it to be identifiable by choosing one representative from each equivalence class. Subject to a modification of this sort, it can be shown (see e.g. [11]) that under very general conditions such models are (a.e.) consistent for the number of components, the mixing proportions, the component parameters, and the density (for data from a finite mixture of the chosen family). It is a common misperception that efficient (approximate) inference is more difficult in these models than in DPMs — to the contrary, we have found that an appropriately constructed MCMC sampler for such a model is nearly identical to an MCMC sampler for a DPM. Further details will be provided elsewhere, since they are beyond the scope of this note.

The rest of the paper is organized as follows. In Section 2, we define the DPM model under consideration. In Section 3, we give an elementary proof of inconsistency for a standard normal DPM. In Section 4, we show (using Hoeffding’s strong law of large numbers for U-statistics) that this example is in fact severely inconsistent, in the sense that the posterior probability of the true number of components goes to 0.

2. SETUP

In this section, we define the Dirichlet process mixture model.

2.1. Dirichlet process mixture model. The DPM model was introduced by Ferguson [3] for the purpose of Bayesian density estimation, and was made practical through the efforts of several authors (see [2] and references therein). We will use $p(\cdot)$ to denote probabilities under the DPM model (as opposed to other probability distributions that will be considered in what follows). The core of the DPM is the so-called Chinese restaurant process (CRP), which defines a certain probability distribution on partitions. Given $n \in \{1, 2, \dots\}$ and $t \in \{1, \dots, n\}$, let $\mathcal{A}_t(n)$ denote the set of all *ordered* partitions (A_1, \dots, A_t) of $\{1, \dots, n\}$ into t nonempty sets. In other words,

$$\mathcal{A}_t(n) = \left\{ (A_1, \dots, A_t) : A_1, \dots, A_t \text{ are disjoint, } \bigcup_{i=1}^t A_i = \{1, \dots, n\}, |A_i| \geq 1 \forall i \right\}.$$

The CRP with concentration parameter $\alpha > 0$ defines a probability mass function on $\mathcal{A}(n) = \bigcup_{t=1}^n \mathcal{A}_t(n)$ by setting

$$p(A) = \frac{\alpha^t}{\alpha^{(n)} t!} \prod_{i=1}^t (|A_i| - 1)!$$

for $A \in \mathcal{A}_t(n)$, where $\alpha^{(n)} = \alpha(\alpha+1) \cdots (\alpha+n-1)$. Note that since t is a function of A , we have $p(A) = p(A, t)$. (It is more common to see this distribution defined in terms of unordered partitions $\{A_1, \dots, A_t\}$, in which case the $t!$ does not appear in the denominator — however, for our purposes it is more convenient to use the distribution on ordered partitions (A_1, \dots, A_t) obtained by uniformly permuting the parts. This does not affect the prior or posterior on t .)

Consider the hierarchical model

$$(2.1) \quad \begin{aligned} p(A, t) &= p(A) = \frac{\alpha^t}{\alpha^{(n)} t!} \prod_{i=1}^t (|A_i| - 1)!, \\ p(\theta_{1:t} \mid A, t) &= \prod_{i=1}^t p(\theta_i), \quad \text{and} \\ p(x_{1:n} \mid \theta_{1:t}, A, t) &= \prod_{i=1}^t \prod_{j \in A_i} p_{\theta_i}(x_j), \end{aligned}$$

where $p(\theta)$ is a prior on component parameters $\theta \in \Theta$, and $\{p_\theta : \theta \in \Theta\}$ is a parametrized family of distributions on $x \in \mathcal{X}$ for the components. Typically, $\mathcal{X} \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^k$ for some d and k . Here, $x_{1:n} = (x_1, \dots, x_n)$ with $x_i \in \mathcal{X}$, and $\theta_{1:t} = (\theta_1, \dots, \theta_t)$ with $\theta_i \in \Theta$. The marginal distribution on $x_{1:n}$ is called a *Dirichlet process mixture (DPM) model*.

The prior on the number of clusters t under this model is $p_n(t) = \sum_{A \in \mathcal{A}_t(n)} p(A, t)$. We use T_n (rather than T) to denote the random variable representing the number of clusters, as a reminder that its distribution depends on n . Note the distinction between the terms “component” and “cluster”: a *component* is part of a mixture distribution, while a *cluster* is the set of (indices of) data points coming from a given component.

Since we are concerned with the posterior distribution $p(T_n = t \mid x_{1:n})$ on the number of clusters, we will be especially interested in the marginal distribution on $(x_{1:n}, t)$, given by

$$(2.2) \quad \begin{aligned} p(x_{1:n}, T_n = t) &= \sum_{A \in \mathcal{A}_t(n)} \int p(x_{1:n}, \theta_{1:t}, A, t) d\theta_{1:t} \\ &= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t \int \left(\prod_{j \in A_i} p_{\theta_i}(x_j) \right) p(\theta_i) d\theta_i \\ &= \sum_{A \in \mathcal{A}_t(n)} p(A) \prod_{i=1}^t m(x_{A_i}) \end{aligned}$$

where for any subset of indices $S \subset \{1, \dots, n\}$, we denote $x_S = (x_j : j \in S)$ and let $m(x_S)$ denote the single-cluster marginal of x_S ,

$$(2.3) \quad m(x_S) = \int \left(\prod_{j \in S} p_\theta(x_j) \right) p(\theta) d\theta.$$

2.2. Specialization to the standard normal case. In this note, for brevity and clarity, we focus on the univariate normal case with unit variance, with a standard normal prior on means — that is, for $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$,

$$p_\theta(x) = \mathcal{N}(x \mid \theta, 1) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}(x - \theta)^2), \quad \text{and} \\ p(\theta) = \mathcal{N}(\theta \mid 0, 1) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}\theta^2).$$

It is a straightforward calculation to show that the single-cluster marginal is then

$$(2.4) \quad m(x_{1:n}) = \frac{1}{\sqrt{n+1}} p_0(x_{1:n}) \exp\left(\frac{1}{2} \frac{1}{n+1} \left(\sum_{j=1}^n x_j\right)^2\right),$$

where $p_0(x_{1:n}) = p_0(x_1) \cdots p_0(x_n)$ (and p_0 is the $\mathcal{N}(0, 1)$ density). When $p_\theta(x)$ and $p(\theta)$ are as above, we refer to the resulting DPM as a *standard normal DPM*.

3. ELEMENTARY EXAMPLE OF INCONSISTENCY

In this section, we prove the following result, exhibiting a simple example in which a DPM is inconsistent for the number of components: the true number of components is 1, but the posterior probability of $T_n = 1$ does not converge to 1. To keep it simple, we set $\alpha = 1$, but the proof extends trivially to any $\alpha > 0$.

Proposition 3.1. *If $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$ i.i.d. then with probability 1, under the standard normal DPM with $\alpha = 1$ as defined above, $p(T_n = 1 \mid X_{1:n})$ does not converge to 1 as $n \rightarrow \infty$.*

Proof. Let $n \in \{2, 3, \dots\}$. Let $x_1, \dots, x_n \in \mathbb{R}$, $A \in \mathcal{A}_2(n)$, and $a_i = |A_i|$ for $i = 1, 2$. Define $s_n = \sum_{j=1}^n x_j$ and $s_{A_i} = \sum_{j \in A_i} x_j$ for $i = 1, 2$. Using Equation 2.4 and noting that $1/(n+1) \leq 1/(n+2) + 1/n^2$, we have

$$\sqrt{n+1} \frac{m(x_{1:n})}{p_0(x_{1:n})} = \exp\left(\frac{1}{2} \frac{s_n^2}{n+1}\right) \leq \exp\left(\frac{1}{2} \frac{s_n^2}{n+2}\right) \exp\left(\frac{1}{2} \frac{s_n^2}{n^2}\right).$$

The second factor equals $\exp(\frac{1}{2} \bar{x}_n^2)$, where $\bar{x}_n = \frac{1}{n} \sum_{j=1}^n x_j$. Writing $s_n/(n+2)$ as a convex combination of $s_{A_1}/(a_1+1)$ and $s_{A_2}/(a_2+1)$, by the convexity of $x \mapsto x^2$ the first factor is less or equal to

$$\exp\left(\frac{1}{2} \frac{s_{A_1}^2}{a_1+1} + \frac{1}{2} \frac{s_{A_2}^2}{a_2+1}\right) = \sqrt{a_1+1} \sqrt{a_2+1} \frac{m(x_{A_1}) m(x_{A_2})}{p_0(x_{1:n})}.$$

Hence,

$$(3.1) \quad \frac{m(x_{1:n})}{m(x_{A_1}) m(x_{A_2})} \leq \frac{\sqrt{a_1+1} \sqrt{a_2+1}}{\sqrt{n+1}} \exp(\tfrac{1}{2} \bar{x}_n^2).$$

Consequently, we have

$$\begin{aligned}
\frac{p(x_{1:n}, T_n = 2)}{p(x_{1:n}, T_n = 1)} &\stackrel{(a)}{=} \sum_{A \in \mathcal{A}_2(n)} n p(A) \frac{m(x_{A_1}) m(x_{A_2})}{m(x_{1:n})} \\
&\stackrel{(b)}{\geq} \sum_{A \in \mathcal{A}_2(n)} n p(A) \frac{\sqrt{n+1}}{\sqrt{|A_1|+1} \sqrt{|A_2|+1}} \exp(-\tfrac{1}{2} \bar{x}_n^2) \\
&\stackrel{(c)}{\geq} \sum_{\substack{A \in \mathcal{A}_2(n): \\ |A_1|=1}} n \frac{(n-2)!}{n! 2!} \frac{\sqrt{n+1}}{\sqrt{2} \sqrt{n}} \exp(-\tfrac{1}{2} \bar{x}_n^2) \\
&\stackrel{(d)}{\geq} \frac{1}{2\sqrt{2}} \exp(-\tfrac{1}{2} \bar{x}_n^2),
\end{aligned}$$

where step (a) follows from applying Equation 2.2 to both numerator and denominator, plus using Equation 2.1 (with $\alpha = 1$) to see that $p(x_{1:n}, T_n = 1) = m(x_{1:n})/n$, step (b) follows from Equation 3.1 above, step (c) follows since all the terms in the sum are nonnegative and $p(A) = (n-2)!/n! 2!$ when $|A_1| = 1$ (by Equation 2.1, with $\alpha = 1$), and step (d) follows since there are n partitions $A \in \mathcal{A}_2(n)$ such that $|A_1| = 1$.

If $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$ i.i.d. then by the law of large numbers, $\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j \rightarrow 0$ almost surely as $n \rightarrow \infty$. Therefore,

$$\begin{aligned}
p(T_n = 1 \mid X_{1:n}) &= \frac{p(X_{1:n}, T_n = 1)}{\sum_{t=1}^{\infty} p(X_{1:n}, T_n = t)} \leq \frac{p(X_{1:n}, T_n = 1)}{p(X_{1:n}, T_n = 1) + p(X_{1:n}, T_n = 2)} \\
&\leq \frac{1}{1 + \frac{1}{2\sqrt{2}} \exp(-\frac{1}{2} \bar{X}_n^2)} \xrightarrow{\text{a.s.}} \frac{1}{1 + \frac{1}{2\sqrt{2}}}.
\end{aligned}$$

Hence, almost surely, $p(T_n = 1 \mid X_{1:n})$ does not converge to 1. \square

Note that the only property of the $\mathcal{N}(0, 1)$ data distribution that we used was $\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j \rightarrow \mathbb{E}X_1 \in \mathbb{R}$. Thus, we could clearly have let X_1, X_2, \dots be i.i.d. from any distribution with finite mean, and still $p(T_n = 1 \mid X_{1:n})$ would not converge to 1.

4. SEVERE INCONSISTENCY

In the previous section, we showed that $p(T_n = 1 \mid X_{1:n})$ does not converge to 1 for a standard normal DPM on standard normal data. In this section, we prove that in fact, it converges to 0. This vividly illustrates that improper use of DPMs can lead to entirely misleading results. The key step in the proof is an application of Hoeffding's strong law of large numbers for U-statistics. The proof generalizes easily to any $\alpha > 0$.

Theorem 4.1. *If $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$ i.i.d. then*

$$p(T_n = 1 \mid X_{1:n}) \xrightarrow{\text{Pr}} 0 \quad \text{as } n \rightarrow \infty$$

under the standard normal DPM with concentration parameter $\alpha = 1$.

Proof. For $t = 1$ and $t = 2$ define

$$R_t(X_{1:n}) = n^{3/2} \frac{p(X_{1:n}, T_n = t)}{p_0(X_{1:n})}.$$

(For general $\alpha > 0$, replace $n^{3/2}$ above by $n^{3/2}\alpha^{(n)}/n!$.) Our method of proof is as follows. We will show that

$$R_2(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{Pr}} \infty$$

(or in other words, for any $B > 0$ we have $\mathbb{P}(R_2(X_{1:n}) > B) \rightarrow 1$ as $n \rightarrow \infty$), and we will show that $R_1(X_{1:n})$ is bounded in probability:

$$R_1(X_{1:n}) = O_P(1)$$

(or in other words, for any $\varepsilon > 0$ there exists $B_\varepsilon > 0$ such that $\mathbb{P}(R_1(X_{1:n}) > B_\varepsilon) \leq \varepsilon$ for all $n \in \{1, 2, \dots\}$). Putting these two together, we will have

$$p(T_n = 1 \mid X_{1:n}) = \frac{p(X_{1:n}, T_n = 1)}{\sum_{t=1}^{\infty} p(X_{1:n}, T_n = t)} \leq \frac{p(X_{1:n}, T_n = 1)}{p(X_{1:n}, T_n = 2)} = \frac{R_1(X_{1:n})}{R_2(X_{1:n})} \xrightarrow[n \rightarrow \infty]{\text{Pr}} 0.$$

First, let's show that $R_2(X_{1:n}) \rightarrow \infty$ in probability. For $S \subset \{1, \dots, n\}$ with $|S| \geq 1$, define $h(x_S)$ by

$$h(x_S) = \frac{m(x_S)}{p_0(x_S)} = \frac{1}{\sqrt{|S|+1}} \exp\left(\frac{1}{2} \frac{1}{|S|+1} \left(\sum_{j \in S} x_j\right)^2\right),$$

where m is the single-cluster marginal as in Equations 2.3 and 2.4. Note that when $1 \leq |S| \leq n-1$, we have $\sqrt{n} h(x_S) \geq 1$. Note also that $\mathbb{E}h(X_S) = 1$ since

$$\mathbb{E}h(X_S) = \int h(x_S) p_0(x_S) dx_S = \int m(x_S) dx_S = 1,$$

using the fact that $m(x_S)$ is a density with respect to Lebesgue measure. For $k \in \{1, \dots, n\}$, define the U-statistics

$$U_k(X_{1:n}) = \frac{1}{\binom{n}{k}} \sum_{|S|=k} h(X_S)$$

where the sum is over all $S \subset \{1, \dots, n\}$ such that $|S| = k$. By Hoeffding's strong law of large numbers for U-statistics [7],

$$U_k(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}h(X_{1:k}) = 1$$

for any $k \in \{1, 2, \dots\}$. Therefore, using Equations 2.1 and 2.2 we have that for any $K \in \{1, 2, \dots\}$ and any $n > K$,

$$\begin{aligned} R_2(X_{1:n}) &= n^{3/2} \sum_{A \in \mathcal{A}_2(n)} p(A) \frac{m(X_{A_1}) m(X_{A_2})}{p_0(X_{1:n})} \\ &= n \sum_{A \in \mathcal{A}_2(n)} p(A) \sqrt{n} h(X_{A_1}) h(X_{A_2}) \\ &\geq n \sum_{A \in \mathcal{A}_2(n)} p(A) h(X_{A_1}) \\ &= n \sum_{k=1}^{n-1} \sum_{|S|=k} \frac{(k-1)!(n-k-1)!}{n! 2!} h(X_S) \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{n-1} \frac{n}{2k(n-k)} \frac{1}{\binom{n}{k}} \sum_{|S|=k} h(X_S) \\
&= \sum_{k=1}^{n-1} \frac{n}{2k(n-k)} U_k(X_{1:n}) \\
&\geq \sum_{k=1}^K \frac{n}{2k(n-k)} U_k(X_{1:n}) \\
&\xrightarrow[n \rightarrow \infty]{\text{a.s.}} \sum_{k=1}^K \frac{1}{2k} = \frac{H_K}{2} > \frac{\log K}{2}
\end{aligned}$$

where H_K is the K^{th} harmonic number, and the last inequality follows from the standard bounds [5] on harmonic numbers: $\log K < H_K \leq \log K + 1$. Hence, for any K ,

$$\liminf_{n \rightarrow \infty} R_2(X_{1:n}) > \frac{\log K}{2} \quad \text{almost surely,}$$

and it follows easily that

$$R_2(X_{1:n}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \infty.$$

Convergence in probability is implied by almost sure convergence.

Now, let's show that $R_1(X_{1:n}) = O_P(1)$. By Equation 2.1, $p(A) = 1/n$ when $A = (\{1, \dots, n\})$. Using this along with Equations 2.2 and 2.4, we have

$$\begin{aligned}
R_1(X_{1:n}) &= n^{3/2} \frac{p(X_{1:n}, T_n = 1)}{p_0(X_{1:n})} = \sqrt{n} \frac{m(X_{1:n})}{p_0(X_{1:n})} \\
&= \frac{\sqrt{n}}{\sqrt{n+1}} \exp\left(\frac{1}{2} \frac{n}{n+1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i\right)^2\right) \leq \exp(Z_n^2/2)
\end{aligned}$$

where $Z_n = (1/\sqrt{n}) \sum_{i=1}^n X_i \sim \mathcal{N}(0, 1)$ for each $n \in \{1, 2, \dots\}$. Since $Z_n = O_P(1)$ then we conclude that $R_1(X_{1:n}) = O_P(1)$. This completes the proof. \square

ACKNOWLEDGMENTS

We would like to thank Stu Geman for raising this question. This research was supported in part by the National Science Foundation under grant DMS-1007593 and the Defense Advanced Research Projects Agency under contract FA8650-11-1-715.

REFERENCES

- [1] M.D. Escobar and M. West, *Bayesian density estimation and inference using mixtures*, Journal of the American Statistical Association **90** (1995), no. 430, 577–588.
- [2] ———, *Computing nonparametric hierarchical models*, Practical Nonparametric and Semiparametric Bayesian Statistics (D. Dey, P. Müller, and D. Sinha, eds.), Springer-Verlag, New York, 1998, pp. 1–22.
- [3] T.S. Ferguson, *Bayesian density estimation by mixtures of normal distributions*, Recent Advances in Statistics (M. H. Rizvi, J. Rustagi, and D. Siegmund, eds.), Academic Press, 1983, pp. 287–302.
- [4] J.K. Ghosh and R.V. Ramamoorthi, *Bayesian Nonparametrics*, Springer-Verlag, New York, 2003.
- [5] R.L. Graham, D.E. Knuth, and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1989.

- [6] P.J. Green and S. Richardson, *Modeling heterogeneity with and without the Dirichlet process*, Scandinavian Journal of Statistics **28** (2001), no. 2, 355–375.
- [7] W. Hoeffding, *The strong law of large numbers for U-statistics*, Institute of Statistics, Univ. of N. Carolina, Mimeograph Series **302** (1961).
- [8] J.P. Huelsenbeck and P. Andolfatto, *Inference of population structure under a Dirichlet process model*, Genetics **175** (2007), no. 4, 1787–1802.
- [9] N. Lartillot and H. Philippe, *A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process*, Molecular Biology and Evolution **21** (2004), no. 6, 1095–1109.
- [10] M. Medvedovic and S. Sivaganesan, *Bayesian infinite mixture model based clustering of gene expression profiles*, Bioinformatics **18** (2002), no. 9, 1194–1206.
- [11] A. Nobile, *Bayesian analysis of finite mixture distributions*, Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1994.
- [12] A. Nobile and A.T. Fearnside, *Bayesian finite mixtures with an unknown number of components: The allocation sampler*, Statistics and Computing **17** (2007), no. 2, 147–162.
- [13] A. Onogi, M. Nurimoto, and M. Morita, *Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods*, BMC Bioinformatics **12** (2011), no. 1, 263.
- [14] E. Otranto and G.M. Gallo, *A nonparametric Bayesian approach to detect the number of regimes in Markov switching models*, Econometric Reviews **21** (2002), no. 4, 477–496.
- [15] S. Richardson and P.J. Green, *On Bayesian analysis of mixtures with an unknown number of components*, Journal of the Royal Statistical Society. Series B **59** (1997), no. 4, 731–792.
- [16] M. West, P. Müller, and M.D. Escobar, *Hierarchical priors and mixture models, with application in regression and density estimation*, Institute of Statistics and Decision Sciences, Duke University, 1994.
- [17] Y. Wu and S. Ghosal, *The L_1 -consistency of Dirichlet mixtures in multivariate Bayesian density estimation*, Journal of Multivariate Analysis **101** (2010), no. 10, 2411–2419.
- [18] E.P. Xing, K.A. Sohn, M.I. Jordan, and Y.W. Teh, *Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture*, Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 1049–1056.