Nonparametric Bayesian Methods for Large Scale Multi-Target Tracking

Emily B. Fox
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139

David S. Choi
MIT Lincoln Laboratory
244 Wood St.
Lexington, MA 02420

Alan S. Willsky
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139

***Abstract*— We consider the problem of data association for multi-target tracking in the presence of an unknown number of targets. For this application, inference in models which place parametric priors on large numbers of targets becomes computationally intractable. As an alternative to parametric models, we explore the utility of nonparametric Bayesian methods, specifically Dirichlet processes, which allow us to put a flexible, data-driven prior on the number of targets present in our observations. Dirichlet processes provide a prior on partitions of the observations among targets whose dynamics are individually described by state space models. These partitions represent the tracks with which the observations are associated. We provide preliminary data association results for the implementation of Dirichlet processes in this scenario.**

I. Introduction

We examine the problem of assigning observations to target tracks when the number of targets present is unknown apriori. In the presence of multiple targets, a sensor will generally collect multiple measurements of each target over time, but lack the inherent ability to associate each measurement with its underlying target. Inference on track assignments given the measurements becomes intractably complex for large numbers of targets when using enumerative or parametric models (the current state-of-the-art). In addition, parametric models can be overly restrictive. As an alternative to parametric model-based association methods, we consider a class of nonparametric Bayesian methods called Dirichlet processes which allow us to put a flexible, data-driven prior on the number of targets present in our observations.

Although termed nonparametric, these Bayesian methods are not free of parameters since they have closed functional forms. Rather, they are processes which learn distributions on function spaces. Therefore, nonparametric Bayesian methods are actually defined by infinitely many parameters. The Dirichlet process is a computationally tractable distribution on probability measures defined over infinite dimensional parameter spaces. In this paper we explore the utility of this distribution as a prior on the unknown number of targets. When combined with an observation likelihood distribution, we obtain a Dirichlet process mixture model, which we will show provides a simple framework for efficiently learning distributions over the space of possible track assignments.

In this paper we first present the motivating application for this work as well as related work in this area. We then describe how the target tracking problem can be formulated as a mixture model and how Dirichlet processes are useful in this context. After describing the model, we present an approach to learning the track assignments and provide results from an example problem. We conclude with a discussion of future work.

II. Background

Data association for target tracking is a challenging problem even when the number of targets being observed is known. When there is little to no prior knowledge about the number of targets, the challenge is compounded. There have been a variety of approaches to solving this problem including brute force methods which enumerate all target track possibilities as well as methods which place parametric prior distributions on the number of targets. There has been a considerable amount of prior research on data association techniques and we will only describe a few of the most relevant methods. For a survey of these methods see [1].

When the number of tracks is known, the joint probabilistic data association filter (JPDAF) performs a time step by time step greedy measurement association. With this formulation, there is no possibility of generating new tracks or terminating old ones. When the number of target tracks is unknown, the multiple hypothesis tracker (MHT) provides a method of data association by enumerating all possible tracks at every time step. The hypothesis space grows exponentially with time, so in practice heuristics must be used to restrict the search space. The approach of Oh, Russell, and Sastry [1] places a Poisson prior on the number of tracks and uses a Markov chain Monte Carlo (MCMC) method to sample the track assignments. Although the Poisson distribution is a valid prior over an arbitrary number of targets, there is a large concentration of probability about the mean of this heavy-tailed distribution. Thus, the Poisson parameter implicitly defines knowledge on the expected number of tracks. Therefore, this parameter must be fine tuned in order to achieve good performance.

By using a Dirichlet process prior we build on the assets of these methods while avoiding the use of heuristics or having to fine-tune parameters. Also, Dirichlet process priors are easily extendible to the more complicated models we consider in the future work section. It is in these extensions that we think the utility of nonparametric Bayesian methods will become apparent.

III. Formulation

We assume we have a data set of noisy range measurements versus time for some unknown number of targets, as shown in Fig.1. Our goal is to assign the measurements to a set of tracks in order to maximize the likelihood of the model. We model the dynamics with the following state space equations,

$$\begin{aligned} &N \sim P_{numtracks}(N) \\ &x_k(t+1) = Ax_k(t) + Bu_k(t), \quad k = 1, \ldots, N \\ &y_k(t) = Cx_k(t) + w_k(t), \qquad k = 1, \ldots, N \\ &\text{detection probability} \quad p_d \end{aligned}$$

where $x_k(t)$ is the state of the k^{th} target at time $t \in \mathbb{Z}_+$, $y_k(t)$ are our observations, $u_k(t) \sim \mathcal{N}(0, Q)$ is process noise on acceleration, and $w_k(t) \sim \mathcal{N}(0, R)$ is measurement noise independent of the process noise. We assume a probability of detection p_d and that the number of targets N is a random variable whose distribution $P_{numtracks}$ may be unknown. We proceed by showing that we can model this system as a finite mixture model when the number of targets N is deterministic and known, and as a Dirichlet process mixture model when N is random with unknown distribution $P_{numtracks}$.

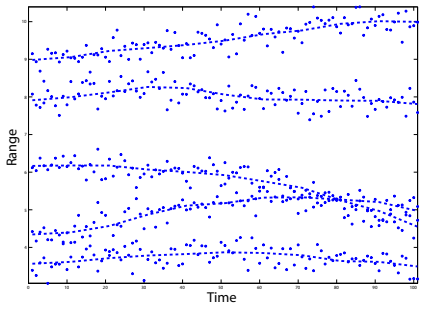


Fig. 1. Noisy range measurements versus time (dots) and true target tracks (dashed lines).

A. Finite Mixture Models

We begin our analysis by assuming that we know there are N targets present. For each target $k = 1, \ldots, N$ we can rewrite the standard state space equations in terms of the initial condition $x_k(0)$ and process noise sequence $\{u_k(1), \ldots, u_k(T)\}$,

$$\begin{aligned} &y_k(t) = CA^t x_k(0) + \sum_{\tau=1}^{t} A^{t-\tau} B u_k(\tau) + w_k(t) \\ &w_k(t) \sim \mathcal{N}(0, R). \end{aligned}$$

It is this set of random variables that uniquely define each track. Therefore, let us define a parameter, θ_k, for each track as follows,

$$\theta_k \triangleq [x_k(0) \quad u_k(1) \quad \ldots \quad u_k(T)]$$

At every time step there may be multiple measurements to assign to the various tracks. If we assume that the j^{th} observation at time t is associated with the k^{th} track, then the observation $y_t^{(j)}$ is distributed according to a Gaussian centered about $\mu_t(\theta_k)$,

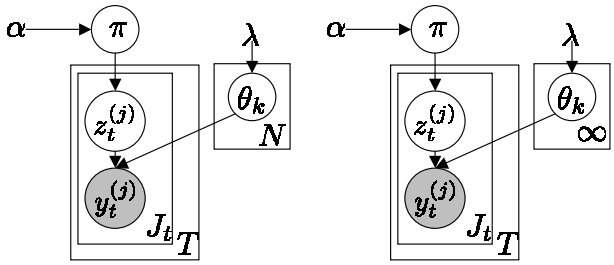
$$f(y_t^{(j)}|\theta_k) = \mathcal{N}(y; \mu_t(\theta_k), R),$$

where $\mu_t(\theta_k) = CA^t x_k(0) + \sum_{\tau=1}^{t} CA^{t-\tau} B u_k(\tau)$.

Let π_k be the prior probability that an observation is generated by the k^{th} target. We can then represent our observations as being generated according to the following mixture of Gaussians,

$$p(y_t^{(j)}|\pi, \theta_1, \ldots, \theta_N) \sim \sum_{k=1}^{N} \pi_k f(y|t, \theta_k).$$

This finite mixture model can be represented by the graphical model shown in Fig.2(a), where each node of the graph represents a random variable in our model and every edge represents the condition dependency between the random variables. The rectangles (or "plates") denote replication of that portion of the graph by the number of times indicated in the lower righthand corner of the rectangle. Here, J_t is the number of measurements at time t and T is the number of time steps in the window of measurements.



$\pi \sim \text{Dir}(\alpha/N, \ldots, \alpha/N)$ (a) $\qquad \pi \sim \text{stick}(\alpha)$ (b)

Fig. 2. (a) Finite mixture graphical model for tracking N targets. (b) Dirichlet process mixture model for tracking an unknown number of targets.

In this model, we have N parameters corresponding to each component of the mixture model. In our scenario, these are the parameters that uniquely define each track. Each observation $y_t^{(j)}$ has an indicator variable, $z_t^{(j)}$, which indexes the parameter associated with the given observation. Therefore, in our scenario, $z_t^{(j)}$ represents the track number. The distribution π determines the mixture weights for each of the conditional densities $f(y|t, \theta_k)$. Thus, π defines a multinomial distribution from which the indicator variables are drawn.

When there is no prior bias towards one component over another, that is to say we do not preference one track over another track, we can define π to be a uniform discrete density. However, there are scenarios in which a uniform density is not appropriate, but we lack prior knowledge on the correct assignment of observations to target tracks. In such situations, we can place a Dirichlet distribution prior on the mixture weights π. A random draw from a Dirichlet distribution is a N-dimensional vector which is constrained to lie within the $(N-1)$-dimensional simplex. Therefore, a Dirichlet distribution defines a distribution over probability measures on

the integers $\{1, \ldots, N\}$. In addition, the Dirichlet distribution is conjugate to the multinomial, which makes it suitable for this application. One can also place a prior distribution $H(\lambda)$ on the parameters θ_k. The generative mixture model can be described by the following equations,

$$\begin{aligned}
&\pi \sim \text{Dir}(\alpha/N, \ldots, \alpha/N) \\
&\theta_k \sim H(\lambda) \\
&z_t^{(j)} \sim \pi \\
&y_t^{(j)} \sim f(y|t, \theta_{z_t^{(j)}}).
\end{aligned}$$

B. Dirichlet Process Mixture Model

In the discussion thus far we have assumed that we know the number of tracks. This begs the question: What if this is unknown apriori and what if we do not want to restrict ourselves to considering a fixed finite number of tracks?

If the number of tracks is allowed to be countably infinite, we need a method of constructing a countably infinite set of mixture weights that satisfy the axioms of probability. In this situation, it is not possible to set π to be a uniform density.

Previously we defined a finite mixture model with a Dirichlet distribution prior on the finite set of mixture weights π (see Fig.2(a)). This can be extended to a countably infinite mixture model by placing a Dirichlet process prior on the countably infinite set of mixture weights (see Fig.2(b)). This gives us what is called a Dirichlet process mixture model [2].

A Dirichlet process defines a distribution over probability measures on potentially infinite parameter spaces Θ. This stochastic process is uniquely defined by a concentration parameter, α, and base measure, H, on the parameter space Θ; we denote it by $DP(\alpha, H)$. A tutorial on Dirichlet processes, including references to seminal work, can be found in [3], [4].

It can be shown that the Dirichlet process actually defines a distribution over discrete probability measures. Namely, w.p.1 a random draw $G \sim DP(\alpha, H)$ is equivalent to $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$, where π_k and θ_k are random. We use the notation δ_{θ_k} to indicate a Dirac delta at θ_k. The weights π_k of this discrete density can be described by the following stick breaking construction. We divide a unit-length stick by the mixture weights π defined over an infinite set of random parameters. The k^{th} mixture weight is a random proportion β_k of the remaining stick after the previous $(k-1)$ weights have been defined,

$$\begin{aligned}
&\beta_k \sim \text{Beta}(1, \alpha) \quad k = 2, 3, \ldots \\
&\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \quad k = 2, 3, \ldots .
\end{aligned}$$

One can easily prove that the axioms of probability are satisfied, namely $\sum_{k=1}^{\infty} \pi_k \overset{\text{a.s.}}{=} 1$. We now have a random probability measure $\pi = \{\pi_k\}_{k=1}^{\infty} \sim \text{stick}(\alpha)$ defined over the positive integers, not just $\{1, \ldots, N\}$. From this construction we see that the parameter α controls the relative proportion of the weights π, and thus controls model complexity in terms of the expected number of components (e.g. tracks).[1]

When the Dirichlet process prior is combined with a likelihood distribution for the observations (as depicted in the graph of Fig.2(b)), we have a Dirichlet process mixture model. The generative mixture model can be described by,

$$\begin{aligned}
&\pi \sim \text{stick}(\alpha) \\
&\theta_k \sim H(\lambda) \\
&z_t^{(j)} \sim \pi \\
&y_t^{(j)} \sim f(y|t, \theta_{z_t^{(j)}}).
\end{aligned}$$

C. Properties of the Dirichlet Process Prior

Because random probability measures drawn from a Dirichlet process are discrete w.p.1, there is a strictly positive probability of associating the same parameter with multiple observations which creates a clustering effect. The data is assigned to a target track based on the parameter with which it is associated.

In addition, there is a reinforcement property that makes it more likely to assign an observation to a track to which other observations have already been assigned. This is described by the predictive distribution of a new track assignment conditioned on all other previous track assignments,

$$p(z_{M+1} = z | z_{1:M}, \alpha, H) = \frac{1}{\alpha + M} (\alpha \delta(z, N+1) + \sum_{k=1}^{N} M_k \delta(z, k)),$$

where M is the total number of observations and M_k are those assigned to the k^{th} track. Here, we use the notation $\delta(z, k)$ to indicate the Kronecker delta. This distribution is the prior distribution on the track assignment of an observation (i.e. the probability of a track assignment when ignoring the likelihood of the observation given that assignment.) We see that the prior probability that the observation was generated from a new, previously unseen track $N+1$ is proportional to α and the prior probability that the observation was generated by an existing track k is proportional to the number of assignments to track k, namely M_k. Therefore, Dirichlet processes favor simpler models. It can be shown that under mild conditions if the data is generated by a finite mixture then the Dirichlet process posterior is guaranteed to converge (in distribution) to that finite set of mixture parameters [5].

IV. Learning

In order to learn the set of track assignments, we use a Markov chain Monte Carlo (MCMC) method, specifically Gibbs sampling. We briefly describe MCMC theory and then present a the Gibbs sampler for our model.

A. Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods [6] are a class of algorithms used to sample from probability distributions

[1] If the value of α is unknown, the model may be augmented with a gamma prior distribution on α, so that the parameter is learned from the data [2].

that are challenging to sample from directly. A Markov chain is constructed whose stationary distribution is the desired density.

$$x^{(n)} \sim q(x|x^{(n-1)}) \qquad n = 1, 2, \ldots$$

After a certain "burn-in" period $\bar{N}$, the state evolution of this chain provides samples from the desired distribution (i.e. $x^{(n)} \sim p(x),\ n > \bar{N}$).

Gibbs sampling is a type of MCMC method that is well suited to state spaces with internal structure. Consider a state space with M states where we wish to draw samples from the joint distribution. With a Gibbs sampler, a sample $x_i^{(n)}$ is drawn from the conditional distribution given the previous set of samples for the other states. We iterate through every state using a specific or random node ordering τ,

$$\begin{aligned}
&x = (x_1, x_2, \ldots, x_M) \\
&\text{for } n = 1 : N_{\text{iter}} \\
&\quad x_i^{(n)} \sim p(x_i|x_{\setminus i}^{(n-1)}) \qquad i = \tau(n) \\
&\quad x_j^{(n)} = x_j^{(n-1)} \qquad\qquad\ \ j \neq \tau(n) \\
&\text{end}
\end{aligned}$$

A node in an undirected graph is conditionally independent of all other nodes given its neighbors. This is equivalent to saying in a directed graph a node is conditionally independent of all other nodes given its Markov blanket, $p(x_i|x_{\mathcal{V}\setminus i}) = p(x_i|x_{MB(i)})$, where the Markov blanket consists of the node's parents, co-parents, and children. Therefore, in the case of sparse graphs, the conditional density from which we are sampling is of much lower dimension than the joint.

B. Gibbs Sampling on Dirichlet Process Mixture Models

In order to infer the set of track assignments $z_t^{(j)}$ in our model, we use a Rao-Blackwellized Gibbs sampler by marginalizing over the infinite set of parameters θ and mixture weights π as in [7]. We map the j^{th} observation at time t to an index i and sample each z_i as follows,

$$\begin{aligned}
&\text{for } n = 1 : N_{\text{iter}} \\
&\quad z_i^{(n)} \sim p(z_i|z_{\setminus i}^{(n-1)}, y, \alpha, \lambda) \qquad i = \tau(n) \\
&\quad z_j^{(n)} = z_i^{(n-1)} \qquad\qquad\qquad\ \ j \neq \tau(n) \\
&\text{end},
\end{aligned}$$

where N_{iter} is the number of iterations of the Gibbs sampler. The conditional density of z_i is proportional to the prior predictive probability of the track assignment times the likelihood of the observation y_i given that track assignment and the other observations and assignments. We can determine the probability of each of the finite set of track assignments as,

$$p(z_i = k|z_{\setminus i}, y, \alpha, \lambda) \propto p(z_i = k|z_{\setminus i}, \alpha) p(y_i|z_i = k, z_{\setminus i}, y_{\setminus i}, \lambda).$$

We note that the conditional dependency between the assignment variables z arises from the marginalization over π.

We can find closed form solutions for the prior and likelihood distributions as follows. The Dirichlet process induces an exchangeable distribution on partitions of the data, so the joint distribution is invariant to the order in which observations are assigned to clusters. Exchangeability implies that we can assume that the i^{th} observation is the last and sample from the predictive distribution of z_i just as we would for z_{M+1},

$$p(z_i|z_{\setminus i}, \alpha) = \frac{1}{\alpha + M - 1}(\alpha\delta(z_i, N+1) + \sum_{j=1}^{N} M_j \delta(z_i, j)).$$

The likelihood distribution is found by analytically marginalizing over θ,

$$\begin{aligned}
p(y_i|z_i = k, z_{\setminus i}, y_{\setminus i}, \lambda) &\propto \int_\Theta \prod_j p(y_j|\theta, z_j) p(\theta|\lambda) d\theta \\
&\propto \int_{\Theta_k} \prod_{j|z_j=k} p(y_j|\theta_k) p(\theta|\lambda) d\theta_k.
\end{aligned}$$

There is a one-to-one mapping between θ_k and X_k, where,

$$X_k \triangleq [x_k(0) \quad x_k(1) \quad \ldots \quad x_k(T)].$$

Therefore, we can equivalently write the likelihood as,

$$\begin{aligned}
p(y_i|z_i = k, z_{\setminus i}, y_{\setminus i}, \lambda) &\propto \int_{X_k} \prod_{j|z_j=k} p(y_j|X_k) p(X_k) dX_k \\
&\propto \int_{\mathcal{X}_k} \prod_{j|z_j=k} p(y_j|x_k(t_j)) \prod_\tau p(x_k(\tau)|x_k(\tau-1)) dX_k,
\end{aligned}$$

where t_j is the time of the j^{th} observation. Let $\{\hat{x}_k, P_k\}$ be the Kalman smoothed state estimate and associated error covariance at t_i generated from $\{y_j|z_j = k, j \neq i\}$, computed by combining the forward filter predictive state statistics with those of the reverse-time filter [8]. The likelihood of the observation y_i is then equivalent to the probability of y_i given the smoothed parameters for its track assignment,

$$\begin{aligned}
p(y_i|z_i = k, z_{\setminus i}, y_{\setminus i}, \lambda) &\propto \int p(y_i|x_k(t_i)) p(x_k(t_i)|\hat{x}_k, P_k) dx_k(t_i) \\
&= \begin{cases} p(y_i|\hat{x}_k, P_k) & k = 1, \ldots, N; \\ p(y_i|P_0) & k = N+1. \end{cases}
\end{aligned}$$

Note that only local changes to the statistics in the smoothing algorithm are needed when observations are reassigned. This allows for an efficient implementation of the Gibbs sampler.

V. Results

We apply the previously described Gibbs sampler to the Dirichlet process mixture model we have presented in order to learn track assignments. To speed up the rate of convergence, we also insert a "switch" step in our sampler as described in [1]. Fig.3 depicts the correct assignment of observations to tracks and the corresponding actual target tracks.

Because MCMC methods provide samples from the posterior distributions after the burn-in period, there are many ways to analyze the results. We examine a few of the statistics resulting from 20,000 iterations of the Gibbs sampler.

In Fig.4(a) we consider the MAP estimate of the track assignments. That is to say, this is the mode of the distribution most often visited by the sampler over 20,000 iterations.

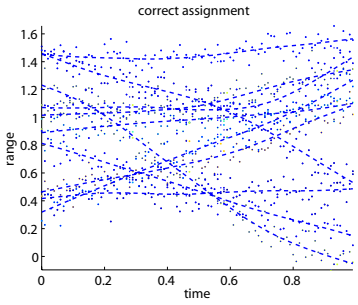


Fig. 3. Correct track assignments.

There are 10 tracks in both the correct assignment and MAP estimate. In addition, though similar, the estimated and true track assignments differ in their target crossing patterns.

Determination of target crossing patterns is an ill-posed problem. In Fig.4(b) we show the seventh most likely assignment. By comparison with the MAP estimate, we see that the Gibbs sampler explores the space of possible crossings. In practice, we probably want to maintain multiple hypotheses of track associations.

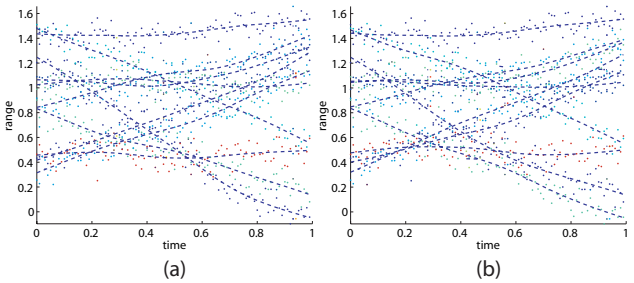


Fig. 4. (a) MAP estimate of track assignments and (b) Seventh most likely track assignments.

We can also consider the problem of track ID. Assume that track ID information for targets 0-9 is known at the beginning of the window of data, as depicted in Fig.5(b). The goal is then to the preserve statistics of track ID at the end of the window by correctly associating a-j with 0-9. Fig.5(a) shows the grid of possible associations. The coloration indicates how likely each association is given our Gibbs samples. The asterisks represent the correct associations. The Munkres algorithm [9] can be applied to find the most likely associations given that each letter a-b can only be assigned to one number in 0-9.

VI. Future Work

We are interested in applying Dirichlet processes to a number of extensions to this model including tracking a single maneuvering target and multiple maneuvering targets as well as groups of targets maneuvering in a coordinated fashion. We are especially interested in the case where the number of maneuver modes is unknown apriori.

For such applications where we are inherently grouping observations which share parameters, we can layer the Dirichlet process priors to create various hierarchical Dirichlet process mixture models [4]. The discussion of these methods is beyond the scope of this paper.

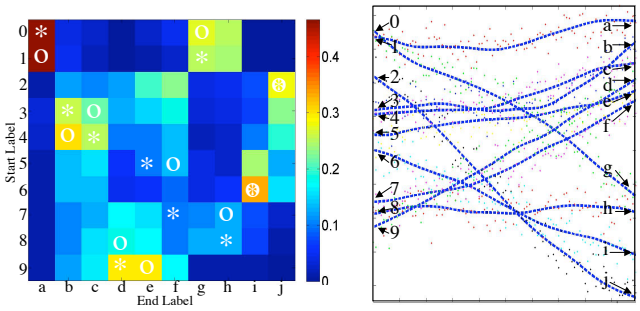


Fig. 5. (a)Probability of start ID to end ID associations (*=correct, o=Munkres). (b) True target tracks and ID associations.

We are also interested in examining a recursive sliding window version of this algorithm for performing on-line data association and tracking.

VII. Conclusion

In conclusion, we have explored a nonparametric Bayesian method for data association in a target tracking application where the number of targets is unknown apriori. Specifically, we have exploited the properties of Dirichlet processes in placing a prior on the number of targets and shown that this provides a flexible and computationally tractable model for learning track associations. We have presented the theoretical background relevant to our modeling choices and have described a learning method using MCMC sampling. Our results indicate the utility of this method in performing data association in the presence of an unknown number of targets.

Acknowledgments

The authors would like to thank Erik Sudderth for his guidance on this project. This research was supported in part by ARO Grant W911NF-05-1-0207, by a MURI funded through AFOSR Grant FA9550-06-1-0324, and by the MIT Lincoln Laboratory DMRI (Decision Modeling Research Initiative). E.B.F. was partially funded by an NDSEG fellowship.

References

[1] S. Oh, S. Russell, and S. Sastry, "Markov chain monte carlo data association for general multiple target tracking problems," in *Proc IEEE Conf on Decision and Control*, December 2004.

[2] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J Amer Stat Assoc*, vol. 90, no. 430, pp. 577–588, 1995.

[3] E. Sudderth, "Graphical models for visual object recognition and tracking," PhD Thesis, MIT, Cambridge,MA, 2006.

[4] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *to appear in Jour Amer Stat Assoc*, 2006.

[5] H. Ishwaran and M. Zarepour, "Bayesian density estimation and inference using mixtures," *Statistica Sinica*, vol. 12, pp. 941–963, 2002.

[6] W. Gilks, S. Richardson, and D. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[7] R. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Jour Comp Graph Stat*, vol. 9, no. 2, pp. 249–265, 2000.

[8] B. Anderson and J. Moore, *Optimal Filtering*. Dover Publications, 2005.

[9] S. Blackman and F. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.