Learning in stochastic dynamical systems is an important and well-studied problem. Recently, Linear Dynamical Systems (LDS) [1], also known as Kalman Filters, have been applied in computer vision to learn dynamic textures [2] from videos of visual phenomena that possess certain stationary properties over time, such as smoke rising, water flowing, etc. Several researchers have improved on the original model, e.g. with Kernel Dynamic Textures [3]. In principle, dynamic textures allow us to learn and generate arbitrarily long realistic sequences of the phenomenon. One impediment to this goal is instability in the underlying dynamics matrix which causes simulated videos to degenerate quickly. This is addressed by our paper at NIPS 2007 [4].

Another drawback of current work in dynamic textures is the inability to represent dynamics that have more structure than the typically used examples such as smoke and water. For example, in a texture with a moving object (such as a swinging pendulum), even a stable LDS will exhibit breakdown and blurring of the object (Fig. 1(B) col. 3). This breakdown results from the LDS’s inherent assumption of a convex probability density: given at least two likely images (such as the pendulum in different positions), their average must be even more likely (resulting in an inconsistent image with the pendulum in multiple positions). Additional examples include flags waving in the wind, time-lapse videos of clouds, and a camera moving past a static texture such as bricks or tiles.

Traditional dynamic texture learning uses PCA on a block Hankel matrix of observations to find a low-dimensional representation of the latent state [2, 4]. It then uses linear regression to predict the state at time $t + 1$ from the state at time $t$, and from the regression weights it constructs an LDS with Gaussian errors (a Kalman filter). In this work, we compare the original and generated state sequences from such a model, and note that the breakdown mentioned above often corresponds to a pattern: the original state sequence lies on a curved submanifold in latent space, while the generated sequence approximately covers the convex hull of this manifold. Typically, the curvature is most obvious in the latent dimensions corresponding to the first few principal components of the Hankel matrix (Fig. 1(B)).

To remedy this problem, we propose learning an LDS with non-Gaussian errors. As a first attempt at such a model, we fit a mixture of $k$ Gaussians to the first few latent dimensions of the state sequence from the training data. We then add this representation of the feasible manifold to the LDS factor graph as a potential (Fig. 1(A)) which intersects with the belief distribution at each time step. The result is that our belief state is always a mixture of Gaussians instead of a single Gaussian. Since the number of components in this mixture grows over time, we limit it to $m$ by sampling.

Experiments are performed on several real-world videos that cannot be handled using existing dynamic texture algorithms, with positive results as shown in Figure 1. One near-term goal is to learn maps of indoor environments using vision data from mobile robots.

References