SCALABLE BAYESIAN REDUCED-ORDER MODELS FOR SIMULATING HIGH-DIMENSIONAL MULTISCALE DYNAMICAL SYSTEMS

PHAEDON-STELIOS KOUTSOURELAKIS† AND ELIAS BILIONIS‡

Abstract. While existing mathematical descriptions can accurately account for phenomena at microscopic scales (e.g., molecular dynamics), these are often high-dimensional, stochastic, and their applicability over macroscopic time scales of physical interest is computationally infeasible or impractical. In complex systems, with limited physical insight on the coherent behavior of their constituents, the only available information is data obtained from simulations of the trajectories of huge numbers of degrees of freedom over microscopic time scales. The analysis of these large amounts of data hinges upon the ability to efficiently extract meaningful latent properties and to discover reduced, predictive descriptions. This paper discusses a Bayesian approach to deriving probabilistic coarse-grained models that simultaneously addresses the problems of identifying appropriate reduced coordinates and the effective dynamics in this lower-dimensional representation. At the core of the models proposed lie simple, low-dimensional dynamical systems which serve as the building blocks of the global model. These approximate the latent generating sources and parametrize the reduced-order dynamics. On their own, each of these simple models would be unable to explain and predict the various complexities encountered in multiscale dynamics of physical interest. Similar to the way one would synthesize opinions from various experts in order to reach a conclusion, we propose probabilistic models that combine the predictions of all these building blocks in order to obtain an integrated model that provides a good global approximation. We discuss parallelizable, online inference and learning algorithms that employ sequential Monte Carlo samplers and scale linearly with the dimensionality of the observed dynamics. We propose a Bayesian adaptive time-integration scheme that utilizes probabilistic predictive estimates and enables rigorous concurrent simulation over macroscopic time scales. The data-driven perspective advocated assimilates computational and experimental data and thus can materialize data-model fusion. It can deal with applications that lack a mathematical description and where only observational data is available. Furthermore, it makes nonintrusive use of existing computational models.

Key words. Bayesian, dimensionality reduction, multiscale, dynamics, sequential inference, time-integration

AMS subject classifications. 65C60, 62M10, 62H25, 62H30, 62F15

DOI. 10.1137/100783790

1. Introduction. The present paper is concerned with the development of probabilistic coarse-grained models for high-dimensional dynamical systems with a view of enabling multiscale simulation. We describe a unified treatment of complex problems described by large systems of deterministic or stochastic ODEs and/or a large number of data streams. Such systems arise frequently in modern multiphysics applications either due to the discrete nature of the system (e.g., molecular dynamics (MD)) or due to discretization of spatio-temporal models (e.g., PDEs):

\[ \frac{dy_i}{dt} = f(y), \quad y \in \mathcal{Y}, \]

Received by the editors January 25, 2010; accepted for publication (in revised form) January 24, 2011; published electronically March 17, 2011. This work was supported by the OSD/AFOSR MURI’09 award to Cornell University on uncertainty quantification.

†School of Civil and Environmental Engineering and Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (pk285@cornell.edu).
‡Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (ib227@cornell.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.
where \( \dim(\mathcal{Y}) >> 1 \) (e.g., \( \mathbb{R}^d, d >> 1 \)). Stochastic versions are also frequently encountered:

\[
\frac{dy_t}{dt} = f(y_t; u_t),
\]

where \( u_t \) is a driving stochastic process (i.e., Wiener process). Uncertainties could also appear in the initial conditions that accompany the aforementioned systems of equations.

Even though the numerical solution of (stochastic) ODEs is a well-studied subject and pertinent computational libraries are quite mature, traditional schemes are impractical or infeasible for systems which are high-dimensional and exhibit a large disparity in scales. This is because most numerical integrators must use time steps of the order of the fastest scales which precludes solutions over long time ranges that are of interest for physical and engineering purposes. In the context of atomistic simulations, practically relevant time scales exceed typical integration steps of \( \sim 1 \text{fs} \) by several orders of magnitude [3]. Furthermore, when numerical solutions of transient PDEs are sought, resolution and accuracy requirements give rise to systems with more than \( 10^9 \) degrees of freedom [108, 24, 135, 77] where the integration time steps are slaved by fast reaction rates or high oscillation frequencies. This impedes their solution and frequently constitutes as computationally infeasible other important tasks such as stability analysis, sensitivity, design, and control.

Multiscale dynamical systems exist independently of the availability of mathematical models. Large numbers of time series appear in financial applications, meteorology, and remote sensing where the phenomena of interest unfold also over a large range of time scales [147, 84]. A wealth of time series data is also available in experimental physics and engineering, which by themselves or in combination with mathematical models can be useful in analyzing underlying phenomena [112, 92, 115] by deriving reduced, predictive descriptions.

Quite frequently the time evolution of all the observables is irrelevant for physical and practical purposes, and the analysis is focused on a reduced set of variables or reaction coordinates \( \hat{y}_t = P(y_t) \) obtained by an appropriate mapping \( P : \mathcal{Y} \to \hat{\mathcal{Y}} \). The goal is then to identify a closed, deterministic or stochastic system of equations with respect to \( \hat{y}_t \), e.g.,

\[
\frac{d\hat{y}_t}{dt} = \hat{f}(\hat{y}_t), \quad \hat{y}_t \in \hat{\mathcal{Y}}.
\]

In the context of equilibrium thermodynamics where ensemble averages with respect to the invariant distribution of \( \hat{y}_t \) are of interest, coarse-graining amounts to free-energy computations [27]. In the nonequilibrium case and when an invariant distribution exists, a general approach for deriving effective dynamics is based on Mori–Zwanzig projections [154, 72, 29, 30, 31, 36]. Other powerful numerical approaches to identify the dynamical behavior with respect to the reduced coordinates include transition path sampling, the transfer operator approach, the nudged elastic band, the string method, Perron cluster analysis, and spectral decompositions [41, 44, 45, 51, 42, 111]. Marked efforts in chemical kinetics have led to an array of computational tools such as computational singular perturbation [104, 105], the intrinsic low-dimensional manifold approach [110, 152], and others [127, 120, 95]. Notable successes in overcoming the time scale dilemma have also been achieved in the context of MD simulations [103, 142, 143, 140] (or Hamiltonian systems in general [119, 106, 134]).
In several problems, physical or mathematical arguments have led analysts to identify a few salient features and their interdependencies that macroscopically describe the behavior of very complex systems consisting of a huge number of individuals/agents/components/degrees of freedom. These variables parametrize a low-dimensional, attracting, invariant, “slow” manifold characterizing the long-term process dynamics [75]. Hence the apparent complexity exhibited in the high-dimensionality and the multiscale character of the original model is a pretext of a much simpler, latent structure that, if revealed, could make the aforementioned analysis tasks much more tractable. The emergence of macroscopic, coherent behavior has been the foundation of coarse-grained dynamic models that have been successful in a wide range of applications. The coarse-grained parametrization and associated model depend on the analysis objectives and particularly on the time scale one wishes to use to make predictions. Modern approaches with general applicability such as the equation-free method [97] or the heterogeneous multiscale method (HeMM) [49] are also based on the availability of reduced coordinates and in the case of HeMM of a macroscopic model which is informed and used in conjunction with the microscale model.

Largely independent of the developments in the fields of computational physics and engineering, the problem of deriving predictive reduced-order models for a large number of time series that potentially exhibit multiple scales has also been addressed in statistics and machine learning communities [148, 56] with applications in network analysis [35], environmetrics [149], sensor network monitoring [150, 124], moving object tracking [4], financial data analysis [5], and computer model emulation [109]. Significant advances have been achieved in modeling [138], forecasting [151], and developing online, scalable algorithms [54, 133, 96, 100, 94, 153] that are frequently based on the discovery of hidden variables that provide insight to the intrinsic structure of streaming data [57, 37, 117, 116, 89].

The present paper proposes a data-driven alternative that is able to automatically obtain coarse-graining of high-dimensional systems without the need of preprocessing and the availability of physical insight. The data is most commonly obtained by simulations of the most reliable, finest-scale (microscopic) model available. This is used to infer a lower-dimensional description that captures the dynamic evolution of the system at a coarser scale (i.e., a macroscopic model). The majority of available techniques address separately the problems of identifying appropriate reduced coordinates and the effective dynamics in this lower-dimensional representation. It is easily understood that the solution of one affects the other. We propose a general framework where these two problems are simultaneously solved and coarse-grained models are built from the ground up. We propose procedures that concurrently infer the macroscopic dynamics and the mapping to the high-dimensional, fine-scale description. As a result, no errors or ambiguity are introduced when the fine-scale model needs to be reinitialized in order to obtain additional simulation data. To that end, we advocate a Bayesian perspective which provides a rigorous foundation for learning from data. It is capable of quantifying inferential uncertainties and, more importantly, uncertainty due to information loss in the coarse-graining process.

We present a Bayesian state-space model where the reduced, coarse-grained dynamics are parametrized by tractable, low-dimensional dynamical models. These can be viewed as experts offering opinions on the evolution of the high-dimensional observables. Each of these modules could represent a single latent regime and would therefore be insufficient by itself to explain the whole system. As is often the case with real experts, their opinions are valid under very specific regimes. We propose therefore a
framework for dynamically synthesizing such models in order to obtain an accurate
global representation that retains its interpretability and computational tractability.

An important contribution of the paper, particularly in view of enabling simula-
tions of multiscale systems, is online inference algorithms based on sequential Monte
Carlo (SMC) that scale linearly with the dimensionality of the observables $d$ (1.1).
These allow the recursive assimilation of data and the recalibration of the coarse-
grained dynamics. The Bayesian framework adopted provides probabilistic predictive
estimates that can be employed in the context of adaptive time-integration. Rather
than determining integration time steps based on traditional accuracy and stability
metrics, we propose using measures of the predictive uncertainty in order to decide
how long into the future the coarse-grained model can be used. When the uncertainty
associated with the predictive estimates exceeds the analyst’s tolerance, the fine-scale
dynamics can be consistently reinitialized in order to obtain additional data that
sequentially update the coarse-grained model.

In agreement with some recently proposed methodologies [97, 98], the data-driven
strategy can seamlessly interact with existing numerical integrators that are well-
understood and reliably implemented in several legacy codes. In addition, it is suitable
for problems where observational/experimental data must be fused with mathematical
descriptions in a rigorous fashion and lead to improved analysis and prediction tools.

The structure of the rest of the paper is as follows. In section 2 we describe
the proposed framework in relation to state-of-the-art approaches in dimensionality
reduction. We provide details of the probabilistic model proposed in the context of
Bayesian state-space models (SSM) in section 2.2. Section 2.3 is devoted to infer-
ence and learning tasks which involve a locally optimal SMC sampler and an online
expectation-maximization scheme. The utilization of the coarse-grained dynamics in
the context of a Bayesian (adaptive) time-integration scheme is discussed in 2.4, and
numerical examples are provided in section 3. Detailed expressions with regard to the
inference and learning algorithms are contained in the appendix.

2. Proposed approach.

2.1. From static-linear to dynamic-nonlinear dimensionality reduction.
The inherent assumption of all multiscale analysis methods is the existence of a lower-
dimensional parametrization of the original system with respect to which the dynam-
elical evolution is more tractable at the scales of interest. In some cases these slow
variables can be identified a priori, and the problem reduces to finding the necessary
closures that will give rise to a consistent dynamical model. In general, however, one
must identify the reduced space $\hat{Y}$ as well as the dynamics within it.

A prominent role in these efforts has been held by methods based on principal
component analysis (PCA). With small differences and depending on the community,
other terms such as proper orthogonal decomposition (POD) or Karhunen–Loève
expansion (KL), empirical orthogonal functions (EOF) have also been used. PCA finds
its roots in the early papers by Pearson [118] and Hotelling [88] and was originally
developed as a static dimensionality reduction technique. It is based on projections
on a reduced basis identified by the leading eigenvectors of the covariance matrix
$C$. In the dynamic case and in the absence of closed form expressions for the actual
covariance matrix, samples of the process $y_t \in \mathbb{R}^d$ at $N$ distinct time instants $t_i$ are
used in order to obtain an estimate of the covariance matrix

$$
C \approx C_N = \frac{1}{N-1} \sum_{i=1}^{N} (y_{t_i} - \mu)(y_{t_i} - \mu)^T,
$$

(2.1)
where $\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$ is the empirical mean. If there is a spectral gap after the first $k$ eigenvalues and $V_K$ is the $d \times K$ matrix whose columns are the $K$ leading normalized eigenvectors of $C_N$, then the reduced-order model is defined with respect, to $\hat{y}_t = V_K^T y_t$. The reduced dynamics can be identified by a Galerkin projection (or a Petrov–Galerkin projection) of the original ODEs in (1.1):

$$\frac{d\hat{y}_t}{dt} = V_K^T f(V_K \hat{y}_t).$$

Hence the reduced space $\hat{Y}$ is approximated by a hyperplane in $Y$ and the projection mapping $P$ linear (Figure 1(a)). While it can be readily shown that the projection adopted is optimal in the mean square sense for stationary Gaussian processes, it is generally not so in cases where non-Gaussian processes or other distortion metrics are examined. The application of PCA-based techniques to high-dimensional, multiscale dynamical systems poses several modeling limitations. Firstly, the reduced space $\hat{Y}$ might not be sufficiently approximated by a hyperplane of dimension $K << d$. Even though this assumption might hold locally, it is unlikely that this will be a good global approximation. Alternatively, the dynamics of the original process might be adequately approximated on a $K$-dimensional hyperplane, but this hyperplane might change in time. Secondly, despite the fact that the projection on the subspace spanned by the leading eigenvectors captures most of the variance of the original process, in cases where this variability is due to the fast modes, there is no guarantee that $\hat{y}_t$ will account for the long-range, slow dynamics which is of primary interest in multiscale systems. Thirdly, the basic assumption in the estimation of the covariance matrix is that the samples $y_t$ are drawn from the same distribution, i.e., that the process $y_t$ is stationary. A lot of multiscale problems, however, involve nonstationary dynamics (e.g., nonequilibrium MD [82, 32]). Hence even if a stationary reduced-order process provides a good, local approximation to $y_t$, it might need to be updated in time. Apart from the aforementioned modeling issues, significant computational difficulties are encountered for high-dimensional systems ($d = \dim(Y) >> 1$), and large datasets ($N >> 1$) as the $K$ leading eigenvectors of large matrices (of dimension proportional to $d$ or $N$) need to be evaluated. This effort must be repeated if more samples become available (i.e., $N$ increases) and an update of the reduced-order model is desirable. Recent efforts have concentrated on developing online versions [145] that circumvent this problem.

![Fig. 1. The phase space is assumed two-dimensional for illustration purposes; i.e., $y_t = (y_{t(1)}, y_{t(2)})$. Each black circle corresponds to a realization $y_{t(i)}$. $P : Y \to \hat{Y}$ is the projection operator from the original high-dimensional space $Y$ to the reduced space $\hat{Y}$.](image-url)
The obvious extension to the linear projections of PCA is nonlinear dimensionality reduction techniques. These have been the subject of intense research in statistics and machine learning in recent years [128, 122, 137, 46, 130, 11, 10] and fairly recently have found their way to computational physics and multiscale dynamical systems (e.g., [34, 102, 114, 62]). They are generally based on calculating eigenvectors of an affinity matrix of a weighted graph. While they circumvent the limiting, linearity assumption of standard PCA, they still assume that the underlying process is stationary (Figure 1(b)). Even though the system’s dynamics might be appropriately tracked on a lower-dimensional subspace for a certain time period, this might not be invariant across the whole time range of interest. The identification of the dynamics on the reduced space $\hat{Y}$ is not as straightforward as in standard PCA, and, in most cases, a deterministic or stochastic model is fit directly to the projected data points [33, 53, 61].

More important, since the inverse mapping $P^{-1}$ from the manifold $\hat{Y}$ to $Y$ is not available analytically, approximations have to be made in order to find pre-images in the data-space [12, 53]. From a computational point of view, the cost of identifying the projection mapping is comparable to standard PCA as an eigenvalue problem on an $N \times N$ matrix that has to be solved. Updating those eigenvalues and the nonlinear projection operator in cases where additional data become available implies a significant computational overhead, although recent efforts [129] attempt to overcome this limitation.

A common characteristic of the aforementioned techniques is that even though the reduced coordinates are learned from a finite amount of simulation data, there is no quantification of the uncertainty associated with these inferences. This is a critical component not only in cases where multiple sets of reduced parameters and coarse-grained models are consistent with the data, but also for assessing errors associated with the analysis and prediction estimates. It is one of the main motivations for adopting a probabilistic approach in this project. Statistical models can naturally deal with stochastic systems that frequently arise in a lot of applications. Most important, perhaps, even in cases where the fine-scale model is deterministic (e.g., (1.1)), a stochastic reduced model provides a better approximation that can simultaneously quantify the uncertainty arising from the information loss that takes place during the coarse-graining process [55, 101].

A more general perspective is offered by latent variable models where the observed data (experimental or computationally generated) is augmented by a set of hidden variables [14]. In the case of high-dimensional, multiscale dynamical systems, the latent model corresponds to a reduced-order process that evolves at scales of practical relevance. Complex distributions over the observables can be expressed in terms of simpler and tractable joint distributions over the expanded variable space. Furthermore, structural characteristics of the original, high-dimensional process $y_t$ can be revealed by interpreting the latent variables as generators of the observables.

In that respect, a general setting is offered by Hidden Markov Models (HMM) [67] or more generally SSM [20, 68, 84]. These assume the existence of an unobserved (latent) process $\hat{y}_t \in \mathbb{R}^K$ described by a (stochastic) ODE:

$$\frac{d\hat{y}_t}{dt} = \hat{f}(\hat{y}_t; w_t) \quad \text{(transition equation)},$$

which gives rise to the observables $y_t \in \mathbb{R}^d$ as

$$y_t = h(\hat{y}_t, v_t) \quad \text{(emission equation)},$$
where \( w_t \) and \( v_t \) are unknown stochastic processes (to be inferred from data) and 
\[
\hat{f} : \mathbb{R}^K \rightarrow \mathbb{R}^K, \ h : \mathbb{R}^K \rightarrow \mathbb{R}^d
\]
are unknown measurable functions. The transition equation defines a prior distribution on the coarse-grained dynamics whereas the emission equation defines the mapping that connects the reduced-order representation with the observable dynamics. The object of Bayesian inference is to learn the unobserved (unknown) model parameters from the observed data. Hence the coarse-grained model and its relation to the observable dynamics are inferred from the data.

The form of (2.3) and (2.4) affords general representations. Linear and nonlinear PCA models arise as special cases by appropriate selection of the functions and random processes appearing in the transition and emission equations. Note, for example, that the transition equation (2.3) for \( \hat{y}_t \) in the case of the PCA-based models reviewed earlier is given by (2.2), and the emission equation (2.4) that relates latent and observed processes is linear, deterministic, and is specified by the matrix of leading eigenvectors \( V_K \).

An extension to HMM is offered by switching-state models [78, 26, 76, 131] which can be thought of as dynamical mixture models [25, 69]. The latent dynamics consist of a discrete process that takes \( M \) values, each corresponding to a distinct dynamical behavior. This can be represented by an \( M \)-dimensional vector \( z_t \) whose entries are zero except for a single one \( m \) which is equal to one and represents the active mode/cluster. Most common, the time-evolution of \( z_t \) is modeled by a first-order stationary Markov process:

\[
(2.5)
\]

\[
z_{t+1} = Tz_t,
\]

where \( T = [T_{m,n}] \) is the transition matrix and \( T_{m,n} = Pr[z_{m,t+1} = 1 \mid z_{n,t} = 1] \). In addition to \( z_t \), \( M \) processes \( x^{(m)}_t \in \mathbb{R}^K \), \( m = 1, \ldots, M \), parametrize the reduced-order dynamics (see also the discussion in section 2.2). Each is activated when \( z_{m,t} = 1 \).

In the linear version (switching linear dynamic system (SLDS)\(^1\)) and conditioned on \( z_{m,t} = 1 \), the observables \( y_t \) arise by a projection from the active \( x^{(m)}_t \) as follows:

\[
(2.6)
\]

\[
y_t = P^{(m)}x^{(m)}_t + v_t, \quad v_t \sim N(0, \Sigma) \text{ independently and identically distributed (i.i.d.),}
\]

where \( P^{(m)} \) are \( d \times K \) matrices (\( K < d \)) and \( \Sigma \) is a positive definite \( d \times d \) matrix. Such models provide a natural, physical interpretation according to which the behavior of the original process \( y_t \), is segmented into \( M \) regimes or clusters, the dynamics of which can be low-dimensional and tractable. From a modeling point of view, the idea of utilizing a mixture of simple models provides great flexibility [18, 16, 132, 74], as it can be theorized that given a large enough number of such components, any type of dynamics can be sufficiently approximated. In practice, however, a large number might be needed, resulting in complex models containing a large number of parameters.

Such mixture models have gained prominence in recent years in the machine learning community. In [17], for example, a dynamic mixture model was used to analyze a huge number of time series, each corresponding to a word in the English vocabulary as they appear in papers in the journal Science. The latent discrete variables represented topics, and each topic implied a distribution on the space of words. As a result, not only a predictive summary (dimensionality reduction) of the high-dimensional observables was achieved but also an insightful deconstruction of the original time series.

\(^1\)These are sometimes referred to as jump-linear or conditional Gaussian models.
was made possible. In fact current research in statistics has focused on infinite mixture models where the number of components can be automatically inferred from the data [136, 23, 13, 59, 60]. In the context of computer simulations of high-dimensional systems, such models have been employed by [58, 86, 84, 83, 85] where maximum likelihood techniques were used to learn the model parameters.

In the next sections we present a novel model that generalizes SLDS. Unlike mixture models which assume that \( y_t \) is the result of a single reduced-order process at each time instant, we propose a partial-membership model (referred to henceforth as partial-membership linear dynamic system (PMLDS)), which allows observables to have fractional memberships in multiple clusters (Figure 2). The latent building blocks are experts [90, 91, 81, 79] which, on their own, provide an incomplete, biased prediction, but when their “opinions” are appropriately synthesized, they can give rise to a highly accurate aggregate model.

From a modeling perspective such an approach has several appealing properties. The integrated coarse-grained model can be interpretable and low-dimensional even for large, multiscale systems, as its expressive ability does not hinge upon the complexity of the individual components but rather is a result of its factorial character [70]. Intricate dynamical behavior can be captured and decomposed in terms of simple building blocks. It is highly suited for problems that lack scale separation and where the evolution of the system is the result of phenomena at a cascade of scales. Each of these scales can be described by a latent process, and the resulting coarse-grained model will account not only for the slow dynamics but also quantify the predictive uncertainty due to the condensed, fast-varying features.

From an algorithmic point of view we present parallelizable, online inference/learning schemes, which can recursively update the estimates produced as more data become available, i.e., if the time horizon \( t \) of the observables \( y_{1:t} \) increases. Unlike some statistical applications where long time series are readily available, in the majority of problems involving computational simulations of high-dimensional, multiscale systems, data is expensive (at least over large time horizons), as they imply calls to the microscopic solvers. The algorithms presented are capable of producing predictive estimates “on the fly” and if additional data are incorporated, they can readily update the model parameters. In addition, such schemes can take advantage of the natural tempering effect of introducing the data sequentially, which can further facilitate the solution of the global estimation problem. More important, perhaps, the updating schemes discussed have linear complexity with respect to the dimensionality \( d \) of the original process \( y_t \).

2.2. PMLDS. We present a hierarchical Bayesian framework which promotes sparsity, interpretability, and efficiency. The framework described can integrate heterogeneous building blocks and allows for physical insight to be introduced on a case-by-case basis. When dealing with high-dimensional molecular ensembles, for example, each of these building blocks might be an (overdamped) Langevin equation with a harmonic potential [19, 84, 87]. It is obvious that such a simplified model would perhaps provide a good approximation under specific, limiting conditions (e.g., at a persistent metastable state) but definitely not across the whole time range of interest. Due to its simple parametrization and computational tractability, it can easily be trained to represent one of the “experts” in the integrated reduced-model. It is known that these models work well under specific regimes, but none of them gives an accurate global description. In the framework proposed, they can, however, be utilized in a way that combines their strengths but also probabilistically quantifies their limitations.
A transient, nonlinear PDE can be resolved into several linear PDEs whose simpler form and parametrization makes them computationally tractable over macroscopic time scales and permits a coarser spatial discretization. Their combination with time-varying characteristics can give rise to an accurate global approximation. Their simplicity and limited range of applicability would preclude their individual use. In the framework proposed, however, these simple models would only serve as good local approximants and their inaccurate predictions would be synthesized into an accurate, global model.

2.2.1. Representations with simple building blocks. We describe a probabilistic, dynamic, continuous-time, generative model which relates a sequence of the observations \( \mathbf{y}_t \in \mathbb{R}^d \) at discrete time instants \( t = 1, 2, \ldots, T \) with a number of hidden processes. The proposed model consists of \( M \) hidden processes \( \mathbf{x}_t^{(m)} \in \mathbb{R}^K, m = 1, \ldots, M (K << d) \), which are assumed to evolve independently of each other and are described by a set of (stochastic) ODEs:

\[
\frac{d\mathbf{x}_t^{(m)}}{dt} = g_m(\mathbf{x}_t^{(m)}, \theta_x^{(m)}), \quad m = 1, \ldots, M.
\]

This equation essentially implies a prior distribution on the space of hidden processes parametrized by a set of (unknown a priori) parameters \( \theta_x^{(m)} \). It should be noted that while the proposed framework allows for any type of process in (2.7), it is desirable that these are simple, in the sense that the parameters \( \theta_x^{(m)} \) are low-dimensional and can be learned swiftly and efficiently. We list some desirable properties of the prior models [132]:

- **Stationarity.** Unless specific prior information is available, it would be unreasonable to impose a time bias on the evolution of any of the reduced dynamics processes. Hence it is important that the models adopted are a priori stationary. Note that the posterior distributions might still exhibit nonstationarity.
- **Correlation decay.** We assume that for any \( m \) and \( t_1, t_2, \) the correlation \( \mathbf{x}_{t_1}^{(m)} \) and \( \mathbf{x}_{t_2}^{(m)} \) should decay monotonically as \( |t_2 - t_1| \) goes to \( +\infty \). This precludes models that do not explicitly account for the time evolution of the latent processes and assume that hidden states are not time-dependent (e.g., static PCA models).
- **Other.** Although this is not necessary, we adopt a continuous time model in the sense of [144] with an analytically available transition density which allows an inference to be carried out seamlessly even in cases where the observables are obtained at nonequidistant times. As a result, the proposed framework can adapt to the granularity of the observables and also provide exact probabilistic predictions at any time resolution.

Even though more complex models can be adopted, we assume here that independent, isotropic Ornstein–Uhlenbeck (OU) processes are used to model the hidden dynamics \( \mathbf{x}_t^{(m)} \) [15]. The OU processes used comply with the aforementioned desiderata. In particular, the following parametrization is employed:

\[
\frac{d\mathbf{x}_t^{(m)}}{dt} = -b_x^{(m)}(\mathbf{x}_t - \mathbf{q}_x^{(m)})dt + (S_x^{(m)})^{1/2}d\mathbf{W}_t^{(m)},
\]

where \( \mathbf{W}_t^{(m)} \) are Wiener processes (independent for each \( m \)), \( b_x^{(m)} > 0, \mathbf{q}_x^{(m)} \in \mathbb{R}^K, \) and \( S_x^{(m)} \) are symmetric, positive definite matrices of dimension \( K \times K \). The aforementioned model has a Gaussian invariant (stationary) distribution \( \mathcal{N}(\mathbf{q}_x^{(m)}, \frac{1}{2b_x^{(m)}}S_x^{(m)}) \).
The transition density denoted by \( p(x_t^{(m)} | x_{t-1}^{(m)}) \) for time separation \( \delta t \) is also a Gaussian \( \mathcal{N}(\mu_{t,\delta t}, \Sigma_{t,\delta t}) \), where
\[
\begin{align*}
\mu_{t,\delta t} &= x_{t-1}^{(m)} - (1 - e^{-k_t^{(m)} \delta t}) (x_{t-1}^{(m)} - q_{x}^{(m)}), \\
\Sigma_{t,\delta t} &= \frac{1-e^{-2k_t^{(m)} \delta t}}{2b_t^{(m)}} \Sigma_x^{(m)}.
\end{align*}
\]

It is not expected that simple processes on their own will provide good approximations to the essential dynamics exhibited in the data \( y_t \). In order to combine the dynamics implied by the \( M \) processes in (2.7), we consider an \( M \)-dimensional process \( z_t \) such that \( \sum_{m=1}^{M} z_{m,t} = 1 \) and \( z_{m,t} > 0 \) for all \( t \) and define an appropriate prior. The coefficients \( z_{m,t} \) express the weight or fractional membership to each process/expert \( x_t^{(m)} \) at time \( t \) [80]. We use the logistic-normal model [17] as a prior for \( z_t \). It is based on a Gaussian process \( \hat{z}_t \) whose dynamics are also prescribed by an isotropic OU process
\[
\begin{align*}
d\hat{z}_t &= -b_z (\hat{z}_t - q_z) dt + S_z^{1/2} dW_t
\end{align*}
\]
and the transformation
\[
\begin{align*}
z_{m,t} &= \frac{e^{\hat{z}_{m,t}} + 1/M}{\sum_{m=1}^{M} e^{\hat{z}_{m,t}} + 1} \forall m, t.
\end{align*}
\]

It is noted that the presence of \( 1/M \) removes the scale invariance and ensures identifiability of the \( \hat{z}_{m,t} \) from \( z_{m,t} \). The invariant and transition densities of \( \hat{z}_t \) are obviously identical to the ones for \( x_t^{(m)} \) with an appropriate substitution of the parameters. The hidden processes \( \{ x_t^{(m)} \}_{m=1}^{M} \) and associated weights \( z_t \) give rise to the observables \( y_t \) as follows (compare with (2.6)):
\[
\begin{align*}
y_t &= \sum_{m=1}^{M} z_{m,t} P^{(m)} x_t^{(m)} + v_t, \quad v_t \sim \mathcal{N}(0, \Sigma) \text{ (i.i.d.)},
\end{align*}
\]
where \( P^{(m)} \) are \( d \times K \) matrices (\( K < d \)) and \( \Sigma \) is a positive definite \( d \times d \) matrix. The aforementioned equation implies a series of linear projections on hyperplanes of dimension \( K \). The dynamics along those hyperplanes are dictated by a priori independent process \( x_t^{(m)} \). It is noted, however, that the reduced dynamics are simultaneously described by all the hidden processes (Figure 2). This is in contrast to PCA methods where a single such projection is considered and mixture PCA models where even though several hidden processes are used, at each time instant it is assumed that a single one is active. Due to the factorial nature of the proposed model, multiple dynamic regimes can be captured by appropriately combining a few latent states. While mixture models (Figure 1(c)) provide a flexible framework, the number of hidden states might be impractically large. As it is pointed out in [70], in order to encode, for example, a time sequence with 30 bits of information, one would need \( k = 2^{30} \) distinct states. It is noted that even though linear projections are implied by (2.12) and Gaussian noise \( v_t \) is used, the resulting model for \( y_t \) is nonlinear and non-Gaussian, as it involves the factorial combination of \( M \) processes \( \{ x_t^{(m)} \}_{m=1}^{M} \) with \( z_t \). It is finally emphasized that (2.8) and (2.10) for the latent processes prescribe just the prior models. The resulting posterior distribution that arises from their combination with the likelihood (section 2.3) will be non-Gaussian. Finally, the prior models will even
be non-Gaussian if prior distributions are used for the model parameters (e.g., \( q_x^{(m)} \) or \( S_x^{(m)} \)).

The parameters \( z_{m,t} \) express the relative importance of the various reduced models or equivalently the degree to which each data point \( y_t \) is associated with each of the \( M \) reduced dynamics \( x_t^{(m)} \). It is important to note that the proposed model allows for time varying weights \( z_{m,t} \) and can therefore account for the possibility of switching between different regimes of dynamics. Figures 3 and 4 depict a simple example \((d = 1)\) which illustrates the flexibility of the proposed approach.

The unknown parameters of the coarse-grained model consist of

- dynamic variables denoted for notational economy by \( \Theta_t \) (i.e., \( \{x_t^{(m)}\}_{m=1}^M, z_t \) for all \( t \)),
- static variables denoted by \( \Theta \) (i.e., \( \theta_x^{(m)} = (b_x^{(m)}, q_x^{(m)}, S_x^{(m)}) \) in (2.8), \( \theta_z = (b_z, q_z, S_z) \) in (2.10), and \( \{P^{(m)}\}_{m=1}^M, \Sigma \) in (2.12)).

### 2.3. Inference and learning.

Inference in the probabilistic graphical model described involves determining the probability distributions associated with the unobserved (hidden) static \( \Theta \) and dynamic parameters \( \Theta_t \) of the model [71]. In the Bayesian setting adopted, this is the posterior distribution of the unknown parameters of the coarse-grained model. Given the observations (computational or experimental) of the original, high-dimensional process \( y_{1:\tau} = \{y_t\}_{t=1}^\tau \), we denote the posterior by \( \pi(\Theta, \Theta_1:\tau) \):

\[
\pi(\Theta, \Theta_1:\tau) = p(\Theta, \Theta_1:\tau | y_{1:\tau}) = \frac{\text{likelihood}}{\text{prior}} = \frac{p(y_{1:\tau} | \Theta, \Theta_1:\tau) p(\Theta_1:\tau)}{p(y_{1:\tau})}
\]

The normalization constant \( p(y_{1:\tau}) \) is not of interest when sampling for \( \Theta \) or \( \Theta_1:\tau \) but can be quite useful for model validation purposes.
According to (2.12), the likelihood is given by

$$
p(y_{1:t} | \Theta, \Theta_1:\tau) = \prod_{t=1}^{\tau} p(y_t | \Theta, \Theta_t).$$

(2.14)

where the densities in the product are described in (2.17). Equation (2.12) defines the likelihood $p(y_t | \Theta, \Theta_t)$, which is basically the weighted product of the likelihoods under each of the hidden processes/experts $x^{(m)}_t$:

$$
p(y_t | \Theta, \Theta_t) = \frac{1}{c(\Theta, \Theta_t)} \prod_{m=1}^{M} p^{w_{m,t}}(y_t | \Theta, \Theta_t),$$

(2.15)

where the normalizing constant $c(\Theta, \Theta_t)$ ensures that the density integrates to one with respect to $y_t$. According to (2.12),

$$
p_m(y_t | \Theta, \Theta_t) \propto \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(y_t - P^{(m)}x^{(m)}_t)^T \Sigma^{-1}(y_t - P^{(m)}x^{(m)}_t) \right\}.
$$

(2.16)

The likelihood can be written in a more compact form as

$$
p(y_t | \Theta, \Theta_t) \propto \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(y_t - W_t X_t)^T \Sigma^{-1}(y_t - W_t X_t) \right\},$$

(2.17)
where
\begin{equation}
X_i^T = \begin{bmatrix} (x_i^{(1)})^T, (x_i^{(2)})^T, \ldots, (x_i^{(M)})^T \end{bmatrix}
\end{equation}
and
\begin{equation}
W_i = \begin{bmatrix} z_{1,i}P^{(1)}, z_{2,i}P^{(2)}, \ldots, z_{M,i}P^{(M)} \end{bmatrix}.
\end{equation}

The first-order Markovian processes adopted for the prior modeling of the dynamic parameters $\Theta_t$ ((2.8), (2.10), (2.9)) imply that
\begin{equation}
p(\Theta_{1:T}, \Theta) = p(\Theta) \prod_{t=1}^{T} p(\Theta_t | \Theta_{t-1}, \Theta),
\end{equation}
where $p(\Theta_1 | \Theta_0, \Theta) = p(\Theta_1 | \Theta) = \nu_0(\Theta_1 | \Theta)$ is the prior on the initial condition which in this work is taken to be the stationary distribution of the underlying OU processes (see discussion in section 2.2.1) and denoted for notational economy by $\nu_0(\cdot | \Theta)$.

The posterior encapsulates uncertainties arising from the potentially stochastic nature of the original process $y_t$, as well as due to the fact that a finite number of observations were used. The difficulty of the problem is that both the dynamic ($\Theta_{1:T}$) and the static parameters ($\Theta$) are unknown. We adopt a hybrid strategy whereby we sample from the full posterior for the dynamic parameters $\Theta_t$ and provide point estimates for the static parameters $\Theta$ (Figure 5). If uniform priors are used for $\Theta$, then the procedure proposed reduces to a maximum likelihood estimation. Nonuniform priors have a regularization effect which can promote the identification of particular features. While the hybrid strategy proposed is common practice in pertinent models [67], in the current framework it is also necessitated by the difficulty in sampling in the high-dimensional state space of $\Theta$ (note that the projection matrices $P^{(m)}$ in particular are of the dimension of the observables $d$ and $d >> 1$) as well as the need for scalability in the context of high-dimensional systems. The static parameters $\Theta$ are estimated by maximizing the log-posterior:
\begin{equation}
L(\Theta) = \log \pi(\Theta | y_{1:T}) = \log \int \frac{\pi(\Theta, \Theta_{1:T} | y_{1:T})}{p(y_{1:T} | \Theta_{1:T})} d\Theta_{1:T}.
\end{equation}

The maximization of $L(\Theta)$ is more complex than a standard optimization task, as it involves integration over the unobserved dynamic variables $\Theta_{1:T}$. While maximization can be accelerated by using gradient-based techniques (e.g., gradient ascent), the dimensionality of $\Theta$ makes such an approach impractical, as it can be extremely difficult to scale the parameter increments. We propose therefore adopting an expectation-maximization framework (EM) which is an iterative, robust scheme that is guaranteed to increase the log-posterior at each iteration [43, 67]. It is based on constructing a series of increasing lower bounds of the log-posterior using auxiliary distributions $q(\Theta_{1:T})$:
\begin{equation}
L(\Theta) = \log \pi(\Theta | y_{1:T}) = \log \int \frac{\pi(\Theta, \Theta_{1:T} | y_{1:T})}{p(y_{1:T} | \Theta_{1:T})} d\Theta_{1:T}
= \log \int q(\Theta_{1:T}) \frac{\pi(\Theta, \Theta_{1:T} | y_{1:T})}{q(\Theta_{1:T})} d\Theta_{1:T}
\geq \log \int q(\Theta_{1:T}) \log \frac{\pi(\Theta, \Theta_{1:T} | y_{1:T})}{q(\Theta_{1:T})} d\Theta_{1:T} \quad \text{(Jensen's inequality)}
= \mathcal{F}(q, \Theta).
\end{equation}
Monte Carlo estimates and require additional iterations to achieve convergence. Filters. Furthermore, potential omission of the smoothing step could increase the variance of the filtering procedures could yield satisfactory results such as ensemble/unscented Kalman and a maximization step (M-step) with respect to $s$. It is obvious that this inequality becomes an equality when in place of the auxiliary distribution. Nevertheless, SMC samplers involve sequential importance sampling, and structure of the posterior and employs the locally optimal importance sampling distribution. In section 2.3.1 we discuss a particle filter that takes advantage of the particular parallelizable framework for nonlinear, non-Gaussian filtering problems whereby the target distribution $q(s)(\Theta_{1:T}) = \pi(\Theta_{1:T} \mid \Theta(s), y_{1:T})$ is represented with a population of $N$ particles $\Theta_{1:T}^{(s,i)}$ and weights $W(s,i)$ such that the expectation in (2.24) can be approximated by

$$E_{q(s)(\Theta_{1:T})} \log \pi(\Theta_{1:T} \mid y_{1:T}) \approx \sum_{i=1}^{N} W(s,i) \log \pi(\Theta_{1:T}^{(s,i)} \mid y_{1:T}).$$

In section 2.3.1 we discuss a particle filter that takes advantage of the particular structure of the posterior and employs the locally optimal importance sampling distribution. Nevertheless, SMC samplers involve sequential importance sampling, and

**Fig. 5. Algorithm for PMLDS.** It is noted that in lieu of the proposed SMC scheme, other approximate filtering procedures could yield satisfactory results such as ensemble/unscented Kalman filters. Furthermore, potential omission of the smoothing step could increase the variance of the Monte Carlo estimates and require additional iterations to achieve convergence.

It is obvious that this inequality becomes an equality when in place of the auxiliary distribution $q(\Theta_{1:T})$ the posterior $\pi(\Theta_{1:T} \mid \Theta, y_{1:T})$ is selected. Given an estimate $\Theta(s)$ at step $s$, this suggests iterating between an expectation step (E-step) whereby we average with respect to $q(s)(\Theta_{1:T}) = \pi(\Theta_{1:T} \mid \Theta(s), y_{1:T})$ to evaluate the lower bound,

$$(2.23) \quad F(s)(q(s), \Theta) = \int q(s)(\Theta_{1:T}) \log \pi(\Theta_{1:T} \mid y_{1:T}) d\Theta_{1:T}$$

and a maximization step (M-step) with respect to $F(s)(q(s), \Theta)$ (and in particular the first part in (2.23) since the second does not depend on $\Theta$),

$$(2.24) \quad \Theta(s+1) = \arg \max_{\Theta} F(s)(q(s), \Theta) = \arg \max_{\Theta} \int q(s)(\Theta_{1:T}) \log \pi(\Theta_{1:T} \mid y_{1:T}) d\Theta_{1:T}$$

As the optimal auxiliary distributions $q(s)(\Theta_{1:T}) = \pi(\Theta_{1:T} \mid \Theta(s), y_{1:T})$ are intractable, we propose employing an SMC or particle filter [47, 39] scheme for estimating the expectations in the M-Step, i.e., (2.24). SMC samplers provide a parallelizable framework for nonlinear, non-Gaussian filtering problems whereby the target distribution $q(s)(\Theta_{1:T}) = \pi(\Theta_{1:T} \mid \Theta(s), y_{1:T})$ is approximated by sampling from $\pi(\Theta_{1:T} \mid y_{1:T})$ and using the auxiliary importance distribution $q(s)(\Theta_{1:T})$ to approximate the expectation in (2.24).
their performance decays with increasing $\tau$ as the dimension of the state space $\Theta_{1:\tau}$ increases even when resampling and rejuvenation mechanisms are employed [8]. Recent efforts based on exponential forgetting have shown that the accuracy of the approximation can be improved (while keeping the number of particles $N$ fixed) by employing smoothing [73] over a fixed-lag in the past [93, 22].

In this paper we make use of an approximate but highly efficient alternative proposed in [7, 8, 9]. This is based on the so-called split-data likelihood (SDL) first discussed in [123], which consists of splitting the observations into blocks (overlapping or nonoverlapping) of length $L < \tau$ and using the pseudolikelihood which arises by assuming that these blocks are independent. It is shown in [8] that this leads to an alternative Kullback–Leibler divergence contrast function and under some regularity conditions that the set of parameters optimizing this contrast function includes the true parameter. Because the size of the blocks is fixed, the degeneracy of particle filters can be averted and the quality of the approximations can be further improved by applying a backward smoothing step over each block [73]. Let $k$ denote the index of the block of length $L$ considered and $\bar{y}_k = y_{(k-1)L+1:kL}$ and $\bar{\Theta}_k = \Theta_{(k-1)L+1:kL}$. If $\tau = rL$, the likelihood is approximated by

$$p(y_{1:\tau} \mid \Theta, \Theta_{1:\tau}) \approx \prod_{k=1}^{r} p(\bar{y}_k \mid \Theta, \bar{\Theta}_k).$$

If $\nu_0(\cdot \mid \Theta)$ is the stationary/invariant prior density of $\Theta$, then for any $k$, $(\Theta, \bar{\Theta}_k, \bar{y}_k)$ are jointly distributed according to

$$\bar{p}(\Theta, \bar{\Theta}_k, \bar{y}_k) = \pi(\Theta)\nu_0(\Theta_{(k-1)L+1} \mid \Theta) p(y_{(k-1)L+1} \mid \Theta_{(k-1)L+1}, \Theta)$$

$$\prod_{t=(k-1)L+2}^{t=kL} p(\Theta_t \mid \Theta_{t-1}, \Theta)p(y_t \mid \Theta_t, \Theta).$$

In a batch EM algorithm using the split-data likelihood and the $k$th block of data, the M-step would involve maximization with respect to $\Theta$ of (see also (2.24))

$$\bar{Q}(\Theta^{(k-1)}, \Theta) = \int \bar{p}(\bar{\Theta}_k \mid \Theta^{(k-1)}, \bar{y}_k) \log \bar{p}(\Theta, \bar{\Theta}_k, \bar{y}_k) \, d\bar{\Theta}_k.$$ 

We utilize an online EM algorithm where the iteration numbers $s$ coincide with the block index $k$ (i.e., $s = k$) which effectively implies that the estimates for $\Theta$ are updated every time a new data block is considered. The E-step is replaced by a stochastic approximation [126, 21], while the M-step is left unchanged. In particular, at iteration $k$ ($s = k$),

(online E-step) $\bar{Q}(\Theta^{(1:k-1)}, \Theta) = (1 - \gamma_k)\bar{Q}(\Theta^{(1:k-2)}, \Theta) + \gamma_k \int \bar{p}(\bar{\Theta}_k \mid \Theta^{(k-1)}, \bar{y}_k) \log \bar{p}(\Theta, \bar{\Theta}_k, \bar{y}_k) \, d\bar{\Theta}_k$

and update the value of the parameters $\Theta$ as

(online M-step) $\Theta^{(k)} = \arg\max_{\Theta} \bar{Q}(\Theta^{(1:k-1)}, \Theta).$

The algorithm relies on a nonincreasing sequence of positive stepsizes $\{\gamma_k\}_{k \geq 0}$ such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$. In this work we adopted $\gamma_k = \frac{1}{k^a}$ with $a = 0.51$. Naturally the integrals above over the hidden dynamic variables $\Theta_k$ are estimated using SMC-based, particulate approximations of $\bar{p}(\Theta_k \mid \Theta^{(k-1)}, \bar{y}_k)$. For small $L$ the
convergence will in general be slow, as the split-block likelihood assumption will be further from the truth. For larger \( L \), convergence is faster, but the performance of the filter decays. For that purpose we also employed a backward smoothing filter over each block using the algorithm described in [73]. The computational cost of the smoothing algorithm is \( O(N^2 L) \). It is finally noted that even though other approximate or exact filtering and smoothing algorithms could be readily employed for this step such as the Kalman filter and its variations [113], these were not explored in this work.

In practice, and in particular for the distributions utilized in the proposed model (e.g., (2.17)), the EM iterations reduce to calculating a set of (multivariate) sufficient statistics \( \Phi \). In particular, instead of the log-posterior lower bound \( \bar{Q}(\Theta^{1:k-1}, \Theta) \) in (2.29) we update the sufficient statistics as follows:

\[
\Phi^{(k)} = (1 - \gamma_k)\Phi^{(k-1)} + \gamma_k \int \bar{p}(\Theta_k \mid \Theta^{(k-1)}, \bar{y}_k) \phi(\Theta_k) \, d\Theta_k,
\]

(2.31)

where \( \int \bar{p}(\Theta_k \mid \Theta^{(k-1)}, \bar{y}_k) \phi(\Theta_k) \, d\Theta_k \) denotes the set of sufficient statistics associated with the block of data \( \bar{y} = y^{(k-1)L+1:kL} \). Specific details are provided in the appendix. It is finally noted that learning tasks in the context of the probabilistic model proposed should also involve identifying the correct structure (e.g., the number of different experts \( M \)). While this problem poses some very challenging issues which are currently the topic of active research in various contexts (e.g., nonparametric methods), this paper is exclusively concerned with parameter learning. In section 3, we discuss Bayesian validation techniques for assessing quantitatively the correct model structure which are computationally feasible due to the efficiency of the proposed algorithms. A summary of the proposed algorithmic steps is contained in Figure 5.

**2.3.1. Locally optimal SMC samplers.** In this section we discuss Monte Carlo approximations of the expectations appearing in (2.31) with respect to the density \( \bar{p}(\Theta_k \mid \Theta, y_{1:L}) = p(\Theta^{(k-1)L+1:kL} \mid \Theta, y^{(k-1)L+1:kL}) \). Note that in order to simplify the notation we consider an arbitrary block of length \( L \) (e.g., \( k = 1 \)) and do not explicitly indicate the iteration number of the EM algorithm. Hence the target density is

\[
(2.32) \quad p(\Theta_{1:L} \mid \Theta, y_{1:L}) = \frac{1}{p(y_{1:L} \mid \Theta)} \nu_0(\Theta_1 \mid \Theta)p(y_1 \mid \Theta, \Theta_1) \prod_{t=2}^{L} p(\Theta_t \mid \Theta_{t-1}, \Theta)p(y_t \mid \Theta_t, \Theta),
\]

where the dynamic variables are \( \Theta_t = (X_t, z_t) \) ((2.18)). Based on earlier discussions, the evolution dynamics of these variables are independent; i.e.,

\[
(2.33) \quad \nu_0(\Theta_1 \mid \Theta) = \nu_0(X_1 \mid \Theta)\nu_0(z_1 \mid \Theta)
\]

and

\[
(2.34) \quad p(\Theta_{t-1} \mid \Theta_{t-1}, \Theta) = p(X_{t-1} \mid X_{t-1}, \Theta)p(z_{t-1} \mid z_{t-1}, \Theta).
\]
Since there is a deterministic relation between $\tilde{z}_t$ and $z_t$ (2.11), we use them interchangeably. In particular we use $\tilde{z}_t$ in the evolution equations since the initial and transition densities are Gaussian (2.10) and $z_t$ in the likelihood densities, as the expressions simplify in (2.12). The initial and transition densities for $X_t$ are also Gaussian. Given that $x_t^{(m)}$ are a priori independent, we have that

$$
\begin{equation}
(2.35)
\begin{align*}
p(X_t | X_{t-1}, \Theta) &= \prod_{m=1}^{M} p(x_t^{(m)} | x_{t-1}^{(m)}, \Theta) \\
&= N(X_t | \mu_t, S_X),
\end{align*}
\end{equation}
$$

where the mean $\mu_t = \mu_t(X_{t-1})$ is given by (2.9) and $S_X = \text{diag}(S_{x,1}, \ldots, S_{x,M})$ (from (2.9) as well).

SMC samplers operate on a sequence of target densities $p(\Theta_{1:t} | y_{1:t}, \Theta)$ which, for any $t$, are approximated by a set of $n$ random samples (or particles) $\{\Theta_t^{(i)}\}_{i=1}^{n}$. These are propagated forward in time using a combination of importance sampling, resampling, and Markov-Chain-Monte-Carlo–based rejuvenation mechanisms [40, 39, 146, 141]. Each of these particles is associated with an importance weight $W^{(i)}$ ($\sum_{i=1}^{n} W^{(i)} = 1$) which is updated sequentially along with the particle locations in order to provide a particulate approximation:

$$
\begin{equation}
(2.36)
p(\Theta_{1:t} | y_{1:t}, \Theta) \approx \sum_{i=1}^{n} W^{(i)} \delta_{\Theta_t^{(i)} | \Theta_{1:t}^{(i)}}(\Theta_{1:t}),
\end{equation}
$$

where $\delta_{\Theta_t^{(i)} | \Theta_{1:t}^{(i)}}(\cdot)$ is the Dirac function centered at $\Theta_t^{(i)}$. Furthermore, for any measurable $\phi(\Theta_{1:t})$ (as in (2.31)) and for all $t$ [38, 28, 22],

$$
\begin{equation}
(2.37)\sum_{i=1}^{n} W^{(i)} g(\Theta_t, \Theta_{1:t}) \rightarrow \int \phi(\Theta_{1:t}) p(\Theta_{1:t} | y_{1:t}, \Theta) d\Theta_{1:t} \quad \text{(almost surely)}.
\end{equation}
$$

The particles are constructed recursively in time using a sequence of importance sampling densities $q_t(\Theta_t | \Theta_{t-1}, y_t, \Theta)$. The importance weights are determined from the fact that

$$
\begin{equation}
(2.38)\quad p(\Theta_{1:t} | y_{1:t}, \Theta) = p(\Theta_{1:t-1} | y_{1:t-1}, \Theta) \frac{p(\Theta_t | \Theta_{t-1}, \Theta)p(y_t | \Theta_t, \Theta)}{p(y_t | y_{1:t-1}, \Theta)}.
\end{equation}
$$

Let $\{W^{(i)}, \Theta_t^{(i)}\}_{i=1}^{N}$ be the particulate approximation of $p(\Theta_{1:t} | y_{1:t-1}, \Theta)$. Note that for $t = 1$ and for the Gaussian initial densities $\nu_0$ of the proposed model, this consists of exact draws and weights $W^{(i)} = \frac{1}{N}$. At time $t$ we proceed as follows [47]:

1. Sample $\Theta_t^{(i)} \sim q_t(\Theta_t | \Theta_{t-1}^{(i)}, y_t, \Theta)$ for all $i$ and set $\Theta_t^{(i)} \leftarrow (\Theta_t^{(i)}; \Theta_t^{(i)})$;
2. Compute incremental weights

$$
\begin{equation}
(2.39)\quad u_t^{(i)} = \frac{p(\Theta_t^{(i)} | \Theta_{t-1}^{(i)}, \Theta)p(y_t | \Theta_t^{(i)}, \Theta)}{q_t(\Theta_t^{(i)} | \Theta_{t-1}^{(i)}, y_t, \Theta)}
\end{equation}
$$

and update the weights
\[ W^{(i)} = \frac{W^{(i)} u_i^{(i)}}{\sum_{j=1}^{N} W^{(i)} u_j^{(i)}}; \]

3. Compute \( ESS = \frac{1}{\sum_{i=1}^{N} (W^{(i)})^2} \), and if \( ESS < ESS_{\text{min}} \), perform multinomial resampling to obtain a new population with equally weighted particles \( (ESS_{\text{min}} = N/2 \) was used in this study). Set \( t \leftarrow t + 1 \), and go to step 1.

It can be easily established [48] that the locally optimal importance sampling density is

\[ q_t^{\text{opt}}(\Theta_t \mid \Theta_{t-1}, y_t, \Theta) = \frac{p(\Theta_t \mid \Theta_{t-1}, \Theta)p(y_t \mid \Theta_t, \Theta)}{\int p(\Theta_t \mid \Theta_{t-1}, \Theta)p(y_t \mid \Theta_t, \Theta) \, d\Theta_t}. \]

In practice, it is usually impossible to sample from \( q_t^{\text{opt}} \) and/or calculate the integral in the denominator. As a result, approximations are used which nevertheless result in nonzero variance in the estimators. In this paper we take advantage of the fact that the transition density of \( X_t \) as well as the likelihood conditioned on \( z_t \) are Gaussians and propose an importance sampling density of the form

\[ q_t(X_t, \hat{z}_t \mid X_{t-1}, \hat{z}_{t-1}, y_t, \Theta) = p(\hat{z}_t \mid \hat{z}_{t-1}, \Theta) \frac{p(X_t \mid X_{t-1}, \Theta)p(y_t \mid X_t, z_t, \Theta)}{\int p(X_t \mid X_{t-1}, \Theta)p(y_t \mid X_t, z_t, \Theta) \, dX_t}. \]

This implies using the prior to draw \( \hat{z}_t \) and the locally optimal distribution (conditioned on \( \hat{z}_t \) or equivalently \( z_t \)) for \( X_t \). The latter will also be a Gaussian whose mean \( \mu_t \) and covariance \( \tilde{S}_X \) can be readily be established (e.g., using Kalman filter formulas):

\[ \tilde{S}_X = (S_X^{-1} + W_t^T \Sigma^{-1} W_t)^{-1}, \]
\[ \mu_t = \tilde{S}_X (S_X^{-1} \mu_t + W_t^T \Sigma^{-1} y_t). \]

As a result, the incremental weights \( u_t \) are given by

\[ u_t = |\tilde{S}_X|^{1/2} \exp \left\{ \frac{1}{2} \mu_t^T \tilde{S}_X^{-1} \mu_t - \frac{1}{2} \mu_t^T S_X^{-1} \mu_t \right\}. \]

2.4. Prediction and Bayesian adaptive time-integration. Bayesian inference results do not include just point estimates but rather samples from the posterior density, at least with respect to the time-varying parameters \( \Theta_t \). The inferred posterior can be readily used to make probabilistic predictions about the future evolution of the high-dimensional, multiscale process \( y_t \). Given observations \( y_{1:t} = \{y_i\}_{i=1}^{t} \), the predictive posterior for the future state of the system \( y_{t+1:t+T} \) over a time horizon \( T \) can be expressed as \(^2\)

\(^2\)When point estimates for \( \Theta \) are used as proposed, the posterior with respect to \( \Theta \) is simply a delta function.
\[
p(y_{\tau+1:T} \mid y_{1:\tau}) = \int p(y_{\tau+1:T} \mid \Theta, \Theta_{\tau+1:T} \mid y_{1:\tau}) \; d\Theta \; d\Theta_{\tau+1:T} \\
= \int \left( \int p(y_{\tau+1:T} \mid \Theta, \Theta_{\tau+1:T}) \; d\Theta_{\tau+1:T} \right) \times p(\Theta, \Theta_{\tau+1:T} \mid y_{1:\tau}) \; d\Theta \\
= \int \left( \int p(y_{\tau+1:T} \mid \Theta, \Theta_{\tau+1:T}) \; d\Theta_{\tau+1:T} \right) \\
\qquad \times p(\Theta, \Theta_{\tau+1:T} \mid y_{1:\tau}) \; d\Theta \\
\tag{2.45}
\]

The integral above can be approximated using Monte Carlo. In particular given the particulate approximation of the posterior \( p(\Theta, \Theta_{1:T} \mid y_{1:T}) \) (which consists of samples of the dynamic variables \( \Theta \), and the MAP estimate of \( \Theta \)), samples from the prior \( p(\Theta_{\tau+1:T} \mid \Theta, \Theta) \) and subsequently the likelihood \( p(\Theta_{\tau+1:T} \mid \Theta, \Theta) \) can readily be drawn. In fact, given that the latter is a multivariate Gaussian, the predictive posterior will consist of a mixture of Gaussians, one for each sample of \( \Theta_{\tau+1:T} \) drawn.

**Fig. 6.** Bayesian adaptive time-integration and data-model fusion illustrated for a two-dimensional flow. The data generated from computational simulations \( y_{1:T} \) and/or experiments \( y_{1:T} \) are sequentially incorporated in the Bayesian model, and the posterior \( p(\Theta, \Theta_{1:T} \mid y_{1:T}) \) over dynamic and static parameters is updated. The predictive posterior \( p(y_{\tau+1:T} \mid y_{1:T}) \) over the time horizon \( T \) is used to efficiently produce probabilistic predictions of the evolution of the high-dimensional process \( y \) in the future. When the uncertainty associated with those predictions exceeds the analysts’ tolerance, the original system is consistently reinitialized, and more data are generated. These are used to update the (predictive) posterior and to produce additional predictive estimates. It is noted that the tolerance in the predictive uncertainty can also be measured with respect to (low-dimensional) observables, which are usually of interest in practical applications.
The important consequence of the Bayesian framework advocated is that predictive estimates are not restricted to point estimates but whole distributions which can readily quantify the predictive uncertainty. This naturally gives rise to a Bayesian, adaptive, time-integration scheme that allows bridging across time scales while providing quantitative, probabilistic estimates of the accuracy of the coarse-grained dynamics (Figure 6). The distribution of (2.45) is used to probabilistically predict the evolution of the system. The time range over which the reduced model is employed does not have to be specified a priori but can be automatically determined by the variance of the predictive posterior (Figure 6). Once this exceeds the allowable tolerance specified by the analyst, the fine-scale process is reinitialized and more data are obtained, which can in turn be used to update the coarse-grained model. It is emphasized that due to the generative character of the model proposed, the reinitialization can be performed consistently based in general on the emission equations (2.12). In contrast to existing techniques such as projective and coarse-projective integration [64, 66, 65, 95, 121] as well as HeMM [50, 107], there is no need to prescribe lifting and restriction operators, and no ambiguity exists with regard to the appropriateness of the reinitialization scheme. Furthermore, the probabilistic coarse-grained model provides quantitative estimates for its predictive ability and automatically identifies the need for more information from the fine-scale model.

3. Numerical experiments. The first example is intended to validate the accuracy of the proposed online EM scheme and utilize a synthetic dataset. The second example uses actual data and illustrates the superiority of the proposed PMLDS model over existing SLDS models. Finally the third example provides an application in the context of multiscale simulations for the time-dependent diffusion equation.

3.1. Synthetic data. We generated data from the proposed model in order to investigate the ability of the inference and learning algorithms discussed. In particular, we considered a mixture of two $M = 2$, one-dimensional OU processes ($K = 1$) as in (2.8) with $(b^{(1)}, q^{(1)}_z, \Sigma^{(1)}) = (0.1, -5.0, 0.2)$ (slow) and $(b^{(2)}, q^{(2)}_z, \Sigma^{(2)}) = (1., +5.0, 2.0)$ (fast). The logistic normal distribution was used to model the weights associated with each of the hidden processes using an isotropic OU process $dz_t = -b_z(z_t - q_z)dt + \Sigma^{1/2}dW_t$ with $b_z = 1.0$, $q_z = [0, 0]^T$, and $\Sigma_z = [10^0, 0_{10}]$. Two $10 \times 1$ projection vectors $P^m, m = 1, 2$, were generated from the prior $N(0, 100I)$ (see appendix), and $(d = 10)$ time series $y_t$ were produced based on (2.12) with idiosyncratic variances $\Sigma = 0.1^2 I$ and time step $\delta t = 1$. The resulting time series exhibit multimodal, non-Gaussian densities as can be seen in Figure 7(a) as well as two distinct time scales, as it can be seen in the autocovariances plotted in Figure 7(b).

Figure 8 depicts the convergence of the proposed online EM scheme to the reference values of $b^{(m)}_z, m = 1, 2$, for various block sizes $L$ and particle populations $N$. Figure 9 depicts the evolution of the log-likelihood per iteration of the EM algorithm. Figure 10(a) depicts the normalized error in the identified $P^{(m)}_q, m = 1, 2$, and idiosyncratic variances $\Sigma$ precoordinate after 20,000 iterations. In all cases the algorithm exhibits good convergence to the reference values.

3.2. Temperature dataset. The goal of this numerical experiment is to illustrate the interpretability of the proposed model and compare with the switching-state linear model discussed in section 2.1 ((2.6)). For that purpose we utilized the temperature data (in degrees Fahrenheit) of the capitals of the 50 states in the USA ($d = 50$). The data was obtained from http://academic.udayton.edu/
Fig. 7. Densities (a) and autocovariances (b) of times series $y_{t,1}$ (black), $y_{t,2}$ (red), and $y_{t,4}$ (blue) (solid lines). With $(-\circ-)$ the densities and autocovariances of the same times series generated using the learned model parameters using the proposed online EM scheme with $L = 200$ and $N = 200$ (see Figures 8 and 10).

Fig. 8. Convergence of $b_x^{(1)}$ (black) and $b_x^{(2)}$ (red) using the online EM algorithm for three different combinations of $L$ and $N$. $(-\circ-)$ corresponds to $L = 100$, $N = 100$, $(-\diamond-)$ to $L = 200$, $N = 200$, and $(-\triangle-)$ to $L = 20$, $N = 1000$.

Fig. 9. Log-likelihood convergence using the online EM algorithm for three different combinations of $L$ and $N$. $(-\circ-)$ corresponds to $L = 100$, $N = 100$, $(-\diamond-)$ to $L = 200$, $N = 200$, and $(-\triangle-)$ to $L = 20$, $N = 1000$. 
Figure 10. Normalized absolute errors (per coordinate) on the identification of the projection vectors $P^{(m)}$, $m = 1, 2$, and idiosyncratic variances $\sigma_j^2$, $j = 1, \ldots, d$, using the proposed online EM scheme with $L = 200$ and $N = 200$.

kissock/http/Weather/citylistUS.htm, and it represents the average daily temperatures from 01/01/1996 until 01/13/2009 (i.e., 5,127 daily observations).

Figure 11 depicts the posterior memberships corresponding to the SLDS and PMLDS models based on a reduced model with two hidden states ($M = 2$) described by one-dimensional OU processes ($K = 1$). The former assumes that at each time instant, the observables $y_t$ arise from a single hidden process. Hence a single entry of $z_t = [z_{1,t}, z_{2,t}, \ldots, z_{M,t}]$ is equal to 1, and the rest are all equal to 0. The top part of Figure 11 shows the posterior mean of $z_{m,t}, m = 1, 2$. As one would expect, the two states correspond to cold-winter (blue) and hot-summer (red) and alternate periodically (roughly the cold-winter state is active between early November until mid-April and the hot-summer state in the remainder of the calendar year). The top part of Figure 12 depicts the corresponding $P^{(m)}$, $m = 1, 2$, where southern states have higher values and northern states lower. Naturally, winter and summer represent the two extremes, but several intermediate states are also present. The proposed partial-membership model can account for those states without increasing the cardinality of the reduced-order dynamics. As can be seen in the bottom part of Figure 11, which depicts the particulate approximation of the posterior of $z_{m,t}, m = 1, 2$, the two hidden states can also be attributed to the two extremes, but the actual temperatures arise by a weighted combination of these two. Naturally during spring-summer months the weight of the “red” state is higher, and during autumn-winter months the weight of the “blue” state takes over. The posterior of the hidden processes $x_t^{(m)}, m = 1, 2$, is depicted in Figure 13.

In order to quantitatively compare the two models we calculated the average, one-step-ahead, predictive log-likelihood $\log p(y_{t+1} \mid y_1:t)$ for all $t \in [0, T)$:

$$p(y_{t+1} \mid y_{1:t}) = \int \log p(y_{t+1} \mid \Theta, \Theta_{t+1} y_{1:t}) p(\Theta, \Theta_{t+1} \mid y_{1:t}) d\Theta d\Theta_{t+1}$$

$$= \int \log p(y_{t+1} \mid \Theta, \Theta_{t+1} y_{1:t}) p(\Theta_{t+1} \mid \Theta, \Theta_{t}) p(\Theta_{t+1} \mid y_{1:t}) d\Theta_{t+1}.$$  \hfill (3.1)

The latter integral is approximated by Monte Carlo using the MAP estimate of $\Theta$ and the particulate approximation of the posterior for the dynamic variables $\Theta_t$. This
Fig. 11. (Top) Posterior mean of $z_{m,t}$, $m = 1, 2$, based on the SLDS, and (bottom) particulate approximation of the posterior of $z_{m,t}$, $m = 1, 2$, PMLDS. The posterior samples at each time step are based on the data up to that point. These samples depend on the estimated values of static parameters $\Theta$ up to that time instant. As seen in the figure, the information provided in the first two years worth of data suggests a different posterior than the one based on more data points. It is important to note that the inference engine operates on the fly; i.e., data in the past are not reused to update the posterior given the current estimates of the static parameters $\Theta$. As more data is accumulated, the algorithm converges to the “correct” $\Theta$ values and the “correct” posterior. Both results were obtained using the previously discussed online EM scheme with $L = 200$ and $N = 100$.

Fig. 12. Maximum posterior estimates of $P^{(m)} \in \mathbb{R}^{50}$, $m = 1, 2$, based on the SLDS (top) and PMLDS (bottom) models. Both results were obtained using the previously discussed online EM scheme with $L = 200$ and $N = 100$.

provides a measure of how well the model generalizes to a novel observation from the same distribution, as the training data and higher values imply a better model. Table 3.1 reports the average values (in bits) plus/minus the standard deviation (over
Fig. 13. Particulate approximation of the posterior of $x_{(m)}^t$, $m = 1, 2$. Results were obtained using the previously discussed online EM scheme with $L = 200$ and $N = 100$.

$t \in (100, T = 5127)$). Similar calculations were carried out for other model cardinalities (i.e., $M, K$), and in all cases the proposed model exhibited superior performance. This superiority becomes more pronounced as $M$ and $K$ increased, which can be attributed to the factorial character of PMLDS.

Table 3.1

<table>
<thead>
<tr>
<th>$M$</th>
<th>$K$</th>
<th>SLDS</th>
<th>PMLDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>$-179.97 \pm 37.31$</td>
<td>$-171.11 \pm 37.29$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$-170.68 \pm 36.95$</td>
<td>$-141.11 \pm 27.82$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>$-176.40 \pm 34.36$</td>
<td>$-143.81 \pm 25.56$</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>$-166.05 \pm 30.57$</td>
<td>$-117.67 \pm 21.15$</td>
</tr>
</tbody>
</table>

3.3. Transient heat equation. We finally apply the proposed analysis scheme to the one-dimensional transient heat equation:

\[
\begin{align*}
\frac{\partial u}{\partial t} &= \frac{\partial}{\partial x} \left( a(x) \frac{\partial u}{\partial x} \right), & x \in [0, 1], \\
u(0, t) &= u(1, t) = 0 & \forall t.
\end{align*}
\]

The spatial domain was discretized with 1,000 finite elements of equal length, and we considered a “rough” conductivity profile shown in Figure 14(a). The conductivity $a(x)$ in each finite element was assumed constant, and its value was drawn independently from a uniform distribution.\(^3\) For $x \in [0, 0.5]$ we used the uniform $U[0.01, 0.1]$ and for $x \in (0.5, 1]$, $U[0.51, 0.6]$. This naturally resulted in the jump observed in Figure 14(a), which as a consequence gave rise to two distinct slow time scales in the solution profile $u(x,t)$ depicted in Figure 14(b). A rough profile of initial conditions was also used (as can be seen Figure 14(b), $t = 0$). In particular at each node $x_i = 0.001i$, $i = 0, \ldots, 1000$, we set $u(x_i,0) = 10x_i(1 - x_i)(1 + 0.1Z_i)$, where $Z_i \sim N(0,1)$ (i.i.d.).

\(^3\)We considered a single realization of the conductivity profile and solved for it as a deterministic problem. The stochastic PDE where $a(x)$ is random is not considered here.
Upon spatial discretization, we obtain a coupled system of ODEs:

\[
\dot{y}_t + Ky_t = 0,
\]

where \( y_t \in \mathbb{R}^{1001} \) represents the solution at the nodes \( x_i \); i.e., \( y_t = [u(x_1, t), u(x_2, t), \ldots, u(x_{1001}, t)]^T \). In contrast to existing approaches for the same diffusion equation (e.g., [2, 1, 125]) we do not exploit mathematical properties of the PDE in specifying the coarse-grained model, but we rely on data. This data is obtained upon temporal discretization of (3.3) where a time step \( \delta t = 0.0001 \) was used.

As a result, at each time step we obtained a vector of observables \( y_t \) of dimension \( d = 1001 \). This data was incorporated in the Bayesian model proposed using two hidden OU process (\( M = 2 \)) of dimension \( K = 2 \) each. In particular we employed the online EM scheme previously discussed over blocks of length \( L = 10 \) time steps and \( N = 100 \) particles. In particular (see also Figure 6)

- data over 20 times steps \( \delta t, y_{1:20} \) (i.e., corresponding to total time \( 20\delta t = 0.002 \)) were ingested by the Bayesian reduced model, and
- the latter was used to predict the evolution of the system over 500 time steps (i.e., total time \( T = 500\delta t = 0.05 \)),
- the original solver of the governing PDE was then reinitialized using the posterior mean estimate of the state of the system \( y_{520} \) and was run for further 20 time steps. Using the additional data obtained \( y_{521:540} \), the Bayesian model was updated, and the process described was repeated.

It is noted that the proposed Bayesian prediction scheme results in a reduction of the number of fine scale integration time steps by a factor of 25 (\( T/20\delta t = 0.05/0.002 \)), leading to a significant acceleration of the simulation process. Figure 15 depicts the posterior estimates of the solution at various time instants. In all cases these approximate very accurately the exact solution, and these estimates improve as more are accumulated. One of the main advantages of the proposed approach is that apart from single-point estimates one can readily obtain credible intervals that quantify predictive uncertainties due to information loss by the use of the reduced-order dynamic model and the finite amount of data used to learn that model. As seen in Figure 15, these envelop the exact solution and become tighter at larger times. As one would expect, when a larger predictive horizon \( T = 0.1 \) (i.e., 1,000 time steps \( \delta t \)) is used, as be seen in Figures 16 and 17, the predictive uncertainty grows. Such a scheme, however, is twice as efficient, leading to a reduction of computational effort.
Fig. 15. Comparison of predictive posterior estimates (posterior mean and 5% and 95% quantiles) with exact solution $u(x,t)$ at various $t$. Dimensionality reduction factor 500. Time acceleration factor 25.

by a factor of 50 (i.e., $T/20\delta t = 0.1/0.002$). Hence if the analyst is willing to tolerate the additional uncertainty, efficiency gains can be achieved. This supports the arguments made previously with regard to an adaptive Bayesian scheme where the level of predictive uncertainty would be specified and the algorithm would automatically

Fig. 16. Comparison of predictive posterior estimates (posterior mean and 5% and 95% quantiles) with exact solution $u(x,t)$ at various $t$. These results were obtained with a prediction horizon $T = 0.1$ ($\delta t = 0.0001$) in contrast to Figure 15 which were obtained for $T = 0.05$. Dimensionality reduction factor 500. Time acceleration factor 50.
revert to the fine scale model in order to obtain more data and improve the predictive estimates.

4. Conclusions. The proposed modeling framework can extract interpretable reduced representations of high-dimensional systems by employing simple, low-dimensional processes. It simultaneously achieves dimensionality reduction and learning of reduced dynamics.

In many ways the proposed approach builds upon successful mixture models (e.g., [58, 86, 84, 83, 85]) which employ/fit low-dimensional processes to dynamic data, assuming that, at any point in time, a single such process gives rise to the high-dimensional observables. The proposed product model offers a useful extension over preexisting mixture models. It can potentially lead to sparser representations, as the different modalities are not equal to the number of hidden states/processes but to the number of ways that these can be be combined. Moreover, it can decompose dynamics, simultaneously exhibiting a cascade of time scales, as each of these scales can be described by a latent process in the coarse-grained model. The Bayesian framework adopted provides a generalization over single-point estimates obtained through maximum-likelihood procedures. It can quantify uncertainties associated with learning from finite amounts of data and produce probabilistic predictive estimates. It can be used to rigorously perform concurrent simulations with the microscopic model without the need of prescribing ad hoc upscaling and downscaling operators.

The inference and learning schemes discussed possess scalability particularly with regard to the large dimension $d$ of the original process. The algorithms proposed imply $O(d)$ order of operations. Furthermore they can dynamically update the coarse-grained models as more data become available. In a typical scenario, the fine-scale model is reinitialized several times in order to obtain additional information about the system’s evolution that is incorporated in the coarse-grained dynamics “on the fly.”

The Bayesian, statistical perspective can readily be extended to the modeling stochastic dynamical systems. This would require generating more than one realization of the original dynamics, which can nevertheless be incorporated in the coarse-grained models using the same online EM scheme. In fact the loss of information that unavoidably takes place during the coarse-graining results in probabilistic reduced-order models even if the original model was deterministic.
A critical question that offers opportunity for future research on the topic relates to structural learning and in particular with regard to the cardinality of the representation, i.e., the number of hidden processes $x^{(m)}_t$ needed (denoted by $M$ in (2.7)). Treating this as a model selection problem as was done in the examples assumes that there is a single, optimal finite-dimensional representation. Current research activities are centered around nonparametric Bayesian priors over infinite combinatorial structures based on the Dirichlet process paradigm and infinite latent features models (e.g., [136, 74, 79]). These offer an alternative perspective by assuming that the number of building blocks is potentially unbounded and that the observables manifest only a sparse subset of those. As a result, the cardinality of the coarse-grained model can be automatically determined from the data. Another aspect that warrants further investigation is prior modeling of the static parameters. Apart from the regularization effect this offers, it can promote the discovery of desirable features, such as slow-varying essential dynamics and sparse factors (e.g., $P^{(m)}$ in (2.12)), which can advance the interpretability of the results and facilitate the inference process.

Finally, a lot of multiscale dynamical systems of interest are parametrized by high-level inputs (e.g., excitations, initial/boundary conditions) that can be used for stability analysis, sensitivity, and design and control purposes [139, 52, 99, 6]. In some cases, experimental/computational data might also be available which have been obtained under various instantiations of these variables. A deficiency of the current model and an important problem for further investigation involves the derivation of coarse-grained models that depend on these parameters. Such dependencies should be introduced in the coarse-scale dynamics. For example, in the case of an input $u_t$ as in (1.2), the hidden dynamics in (2.7) must be

\[
d\frac{dx^{(m)}_t}{dt} = g_m(x^{(m)}_t; u_t, \theta^{(m)}_x), \quad m = 1, \ldots, M.
\]

This will lead to the discovery of persistent macroscopic features that govern the behavior of the system and to quantification of their dependence on variations to the aforementioned input parameters.

Appendix. This appendix discusses the sufficient statistics and update equations for the static parameters $\Theta$ used in the probabilistic model proposed. In the first section we discuss parameters appearing in the reduced-order dynamics models and in the second those appearing in the likelihood.

Sufficient statistics for parameters appearing in the prior. As discussed in section 2.2.1, independent, isotropic OU processes are used as prior models for the latent, coarse-grained dynamics $x^{(m)}_t \in \mathbb{R}^K$ as well the process $z_t \in \mathbb{R}^M$ that models the fractional memberships to each process $m$. We therefore discuss the essential elements for the online EM computations described in section 2.3 [7, 8, 9] for a general isotropic OU process in $\mathbb{R}^n$ of the form

\[
dx_t = -b(x_t - q)dt + S^{1/2}dW_t.
\]

It is of interest to determine the parameters $\theta = (b, q, S)$. Let $\pi(\theta)$ denote the prior on $\theta$. The readers can adjust the expressions below to any $x^{(m)}_t$ or $z_t$ since independent priors were used. Note that the stationary distribution of $x_t$ is a Gaussian:

\[
\nu_0(x) = \mathcal{N}(x \mid q, C = \frac{1}{2b}S),
\]
and the transition density $p(x_t \mid x_{t-1})$, assuming that equidistant time instants with time step $\delta t$ are considered, is given by

$$
(4.4) \quad p(x_t \mid x_{t-1}) = N(x_t \mid \mu_{st}(x_{t-1}), S_{st}),
$$

where

$$
(4.5) \quad \mu_{st}(x_{t-1}) = x_{t-1} - (1 - e^{-b\delta t})(x_{t-1} - q)
$$

and

$$
(4.6) \quad S_{st} = \frac{1 - e^{-2b\delta t}}{2b} S.
$$

Given a block of length $L$ with observables $y_{1:L}$ and according to (2.27) and (2.28), we have that

$$
\hat{Q}(\Theta^{(k-1)}, \theta) = \langle -\frac{1}{2} \log |C| - \frac{1}{2}(x_1 - q)^T C^{-1}(x_1 - q) + \sum_{t=2}^L -\frac{1}{2} \log |S_{st}| - \frac{1}{2}(x_t - q)^T C^{-1}(x_t - q) \rangle + \log \pi(\theta),
$$

where the brackets $\langle . \rangle$ imply expectation with respect to $\hat{p}(x_{1:L} \mid \Theta^{(k-1)}, y_{1:L})$ as in (2.28). In order to maximize $\hat{Q}(\Theta^{(k-1)}, \Theta)$ as in (2.30), one needs to solve the system of equations arising from $\frac{\partial \hat{Q}(\Theta^{(k-1)}, \Theta)}{\partial \theta} = 0$. These equations with respect to $\theta$ are solved with fixed point iterations. They depend on the following seven sufficient statistics $\Phi = \{ \Phi_j \}_{j=1}^7$:

$$
\Phi_1 = \langle x_1 \rangle,
\Phi_2 = \langle x_1 x_1^T \rangle,
\Phi_3 = \langle \sum_{t=2}^L x_{t-1} \rangle,
\Phi_4 = \langle \sum_{t=2}^L x_t - x_{t-1} \rangle,
\Phi_5 = \langle \sum_{t=2}^L x_{t-1} x_{t-1}^T \rangle,
\Phi_6 = \langle \sum_{t=2}^L (x_t - x_{t-1}) x_{t-1}^T \rangle,
\Phi_7 = \langle \sum_{t=2}^L (x_t - x_{t-1})(x_t - x_{t-1})^T \rangle.
$$

**Sufficient statistics for parameters appearing in the likelihood.** The process is a bit more involved in the case of the parameters appearing in the likelihood equation (2.14), i.e., the projection matrices $\{ P^{(m)} \}_{m=1}^M$ of dimension $d \times K$ and the covariance $\Sigma$ which is a (positive definite) matrix of $d \times d$. In order to retain scalability in high-dimensional problems (i.e., $d >> 1$) we assume a diagonal form of $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$, which implies learning $d$ parameters rather than $d(d+1)/2$. 
Denoting now by \( \theta = ((P^{(m)})^{M}_{m=1}, \{\sigma_j^2\}_{j=1}^d), \pi(\theta) \) the prior, and according to (2.27) and (2.28) we have that

\[
\hat{q}(\theta^{(k-1)}, \theta) = \sum_{t=1}^L -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (y_t - W_t X_t)^T \Sigma^{-1} (y_t - W_t X_t) + \log \pi(\theta).
\]

Differentiation with respect to \( P^{(m)} \) reveals that the stationary point must satisfy

\[
A^{(m)} = \sum_{n=1}^M P^{(n)} B^{(n,m)},
\]

where the sufficient statistics are

\[
A^{(m)} = \left\langle \sum_{i=1}^L z_{t,m} y_t (x_i^{(m)})^T \right\rangle, \quad m = 1, 2, \ldots, M,
\]

and

\[
B^{(n,m)} = \left\langle \sum_{i=1}^L z_{t,n} z_{t,m} x_i^{(n)} (x_i^{(m)})^T \right\rangle.
\]

In the absence of a prior \( \pi(\theta) \) and if \( P_j^{(m)} \) and \( A_j^{(m)} \) represent the \( j \)th rows (\( j = 1, \ldots, d \)) of the matrices \( P^{(m)} \) and \( A^{(m)} \), respectively, then (4.10) implies

\[
\begin{bmatrix}
A_j^{(1)} & A_j^{(2)} & \ldots & A_j^{(M)} \\
A_j:\{1\times K M\}
\end{bmatrix} = \begin{bmatrix}
P_j^{(1)} & P_j^{(2)} & \ldots & P_j^{(M)} \\
P_j:\{1\times K M\}
\end{bmatrix}
\begin{bmatrix}
B^{(1,1)} & B^{(1,2)} & \ldots & B^{(1,M)} \\
B^{(2,1)} & B^{(2,2)} & \ldots & B^{(2,M)} \\
\vdots & \vdots & \ddots & \vdots \\
B^{(M,1)} & B^{(M,2)} & \ldots & B^{(M,M)} \\
B:\{M K \times M K\}
\end{bmatrix}.
\]

This leads to the following update equations for \( P_j^{(m)} \) for all \( j, m \):

\[
P_j = A_j B^{-1}.
\]

Note that the matrix \( B \) to be inverted is independent of the dimension of the observables \( d \) (\( d >> 1 \)) and that the inversion needs to be carried out once for all \( j = 1, \ldots, d \). Hence the scaling of the update equations for \( P^{(m)} \) is \( O(d) \), i.e., linear with respect to the dimensionality of the original system.

Furthermore, in the absence of a prior \( \pi(\theta) \), differentiation with respect to \( \sigma_j^{-2} \) (\( j = 1, \ldots, d \)) leads to the following update equation:

\[
L \sigma_j^2 = \sum_{t=1}^L y_{t,j}^2 - 2A_j P_j^T + P_j B P_j^T.
\]

In summary the sufficient statistics needed are the ones in (4.10) and (4.13).
In the numerical examples in this paper a diffuse Gaussian prior was used for $P^{(m)}$ with variance 100 for each of the entries of the matrix. This leads to the addition of the term $1/100$ in the diagonal elements of the $B$ in (4.13). No priors were used for $\sigma_j^2$.

REFERENCES


