Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Discriminative semi-supervised learning of dynamical systems for motion estimation

Minyoung Kim

Department of Electronic & Information Engineering, Seoul National University of Science & Technology, Seoul 139-743, Republic of Korea

ARTICLE INFO

Available online 9 December 2010

Keywords:
Semi-supervised learning
Linear dynamical systems
Motion estimation
Discriminative learning

ABSTRACT

We introduce novel discriminative semi-supervised learning algorithms for dynamical systems, and apply them to the problem of 3D human motion estimation. Our recent work on discriminative learning of dynamical systems has been proven to achieve superior performance than traditional generative learning approaches. However, one of the main issues of learning the dynamical systems is to gather labeled output sequences which are typically obtained from precise motion capture tools, hence expensive. In this paper we utilize a large amount of unlabeled (input) video data to improve the prediction performance of the dynamical systems significantly. We suggest two discriminative semi-supervised learning approaches that extend the well-known algorithms in static domains to the sequential, real-valued multivariate output domains: (i) *self-training* which we derive as coordinate ascent optimization of a proper discriminative objective over both model parameters and the unlabeled state sequences, (ii) *minimum entropy* approach which maximally reduces the model's uncertainty in state prediction for unlabeled data points. These approaches are shown to achieve significant improvement against the traditional generative semi-supervised learning methods. We demonstrate the benefits of our approaches on the 3D human motion estimation problems.

1. Introduction

We consider the problem of predicting a real-valued multivariate state sequence that undergoes certain dynamics. A typical example is the human motion estimation in computer vision where we want to estimate the sequence of human body joint angles from video measurements (a sequence of image frames). Formally, we denote the state sequence to be predicted as $\mathbf{Y}=\mathbf{y}_1\cdots\mathbf{y}_T$, and the measurement sequence of image frames as $\mathbf{X}=\mathbf{x}_1\cdots\mathbf{x}_T$. We assume that the state at time t is d-dim real-valued multivariate vector $\mathbf{y}_t\in\mathbb{R}^d$ (e.g., body joint angles in case of human motion), and the measurement $\mathbf{x}_t\in\mathbb{R}^p$ is the p-dim feature vector at time t (e.g., image features such as the shape descriptors extracted from the silhouette image at video frame t).

Capturing the underlying dynamics in the state sequence is often achieved by adopting a probabilistic model $P(\mathbf{X},\mathbf{Y})$, where the model has dynamical components for the state variables. One of the most popular models in this realm is the state-space model such as the linear dynamical system (LDS). In this paper we focus on the LDS model despite the potential mismatch between its representational capability and complex, nonlinear human motion in real world. However. the linearity assumption of LDS can be easily extended to nonlinear models via nonlinear feature mapping [1]. In addition, the piecewise combination of LDS, often referred to as switching LDS [2], is a widely used technique to enlarge the modeling capacity of LDS. In the pattern recognition community, the use of LDS is also widespread [3–6], in speech, multi-resolution images, and tracking, to name a few.

Since the probabilistic LDS model can serve as the state predictor for a given input $\mathbf{X}$, namely $\mathbf{Y}=\mathbf{F}(\mathbf{X})=\text{argmax}_{\mathbf{Y}}P(\mathbf{Y}|\mathbf{X})$, how to learn the dynamic models is critical for the prediction performance. Unlike the traditional unsupervised learning approach that treats the states $\mathbf{Y}$ as nuisance latent variables and maximizes the measurement likelihood $P(\mathbf{X})$ [7,8], the increased availability of the precise motion capture tools enabled the supervised learning algorithms that can make use of the labeled dataset $\mathcal{L}=\{(\mathbf{X}^i,\mathbf{Y}^i)\}_{i=1}^n$ of possibly different lengths $\{T_i\}_{i=1}^n$.

In the supervised setting, the standard maximum likelihood (ML) learning that optimizes the joint log-likelihood $\sum_{i=1}^n \log P(\mathbf{X}^i,\mathbf{Y}^i)$ can represent a joint density $P(\mathbf{X},\mathbf{Y})$ effectively. However, the ML learning is not directly compatible with the ultimate goal of the state prediction. Rather, the discriminative framework that focuses on the state conditional density $P(\mathbf{Y}|\mathbf{X})$ has recently emerged in the computer vision and related areas [9–11,1,12]. It often outperforms the ML learning especially when the model structure is incorrect and the number of training data points is not small.

In particular, the discriminative learning algorithms for LDS suggested by our earlier work [1,12], more specifically, the conditional maximum likelihood (CML) that optimizes $\sum_{i=1}^n \log P(\mathbf{Y}^i|\mathbf{X}^i)$ and the

E-mail address: mikim21@gmail.com

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.12.002

slicewise CML (SCML) that maximizes $\sum_{i=1}^{n}(1/T_i)\log P(\mathbf{y}_t^i|\mathbf{X}^i)$ have shown outstanding prediction performance in the human motion estimation problems. These approaches have intimate analogy with the discriminative training of hidden Markov models (HMMs) [13] and the conditional random fields (CRFs) [14] which are successful for the structured output *classification* problem that takes *discrete* states $\mathbf{y}_t$ instead of the real vectors. For real-valued multivariate states, the CML and the SCML learning of LDS are especially beneficial for circumventing the difficult density integrability issue that can arise when extending CRFs to a real-valued state domain.

Despite their success, the discriminative learning algorithms (CML and SCML) basically assume a supervised setup, requiring expensive label (i.e., $\mathbf{Y}$) collection stage. In this paper, we extend these discriminative learning algorithms to the semi-supervised setting so that we can improve their prediction performance using a large amount of unlabeled sequences (i.e., $\mathbf{X}$ only) which are obtainable with little cost. Formally, we let our training data composed of m sequences: n labeled sequences $\mathcal{L}=\{(\mathbf{X}^i,\mathbf{Y}^i)\}_{i=1}^{n}$, and $(m-n)$ unlabeled sequences $\mathcal{U}=\{\mathbf{X}^j\}_{j=n+1}^{m}$.

In the general semi-supervised learning framework (e.g., [15]), the unlabeled data are often utilized for restricting the model hypothesis space, facilitating regularization framework. More specifically, the semi-supervised learning can be formulated as the following optimization:

$$\max_{\theta} O(\theta,\mathcal{L})-\lambda\cdot R(\theta,\mathcal{U}), \tag{1}$$

where θ is the (LDS) model parameters, $O(\theta,\mathcal{L})$ is the *objective* for the labeled data typically employed by the fully supervised learning, $R(\theta,\mathcal{U})$ is the regularization term acting on the unlabeled data, and λ (≥ 0) is the trade-off constant that balances the objective against the regularization. In this framework the unlabeled data are exploited to reduce the model (hypothesis) space, and hence the learned model has higher chance to yield better generalization performance according to the structural risk minimization theory [16,15].

The fairly standard expectation maximization (EM) based semi-supervised learning that maximizes $\sum_{i\in\mathcal{L}}\log P(\mathbf{X}^i,\mathbf{Y}^i;\theta)+\lambda\sum_{j\in\mathcal{U}}\log P(\mathbf{X}^j;\theta)$ can fall into the above general framework as it employs the negative marginal likelihood on $\mathcal{U}$ as regularization (i.e., $R(\theta,\mathcal{U})=-\sum_{j\in\mathcal{U}}\log P(\mathbf{X}^j;\theta)$). In this way, the EM approach favors and restricts the models θ that have higher marginal likelihoods on the unlabeled data. Thus this results in purely generative models. On the other hand, to have more predictive and discriminative models, one can adopt the CML or SCML learning objective for $O(\theta,\mathcal{L})$. In addition to this, we suggest two discriminative approaches for the regularization part that replace the generative marginal likelihood term with discriminative ones.

The first approach is the *self-training* which first learns the model from the labeled data, and iteratively refines the model by predicting and bootstrapping the labels of the unlabeled data based on the current model through supervised learning. The basic assumption of self-training is that the model's prediction on the unlabeled data points, at least some of them, tends to be correct. The self-training is traditionally framed in the *static* (non-dynamic) semi-supervised setting, and this paper is the first to apply it to a dynamical state prediction task. We also show that the self-training can be viewed as coordinate ascent optimization of a proper discriminative objective over both model parameters (θ) and the unlabeled state sequences ($\{\mathbf{Y}^j\}_{j\in\mathcal{U}}$). The second approach is based on the *minimum entropy* principle suggested by [17]. By taking the entropies of the conditional densities $P(\mathbf{Y}|\mathbf{X}^j)$ on the unlabeled data as the regularization terms, we maximally reduce the model's uncertainty in state prediction for unlabeled data points. The minimum entropy approach was previously shown to yield significantly improved prediction performance in the static classification [17] and the discrete-state structured output classification problems for CRFs [18,19]. In this paper, we extend it to the dynamic continuous-state regression setting of the Gaussian random fields, and derive the gradient evaluation using the nested forward/backward recursion for the semi-supervised learning objectives.

For these two approaches, we show that incorporating the unlabeled data can significantly improve the prediction performance of the LDS models compared to those learned with the labeled data only as well as generative semi-supervised approaches. The rest of the paper is organized as follows: After reviewing some related work in the following, the LDS model is briefly described in the next section with the review of our earlier work on the discriminative learning of LDS in the supervised setting. Section 3 proposes two discriminative semi-supervised learning algorithms. We demonstrate the benefits of our approaches on both synthetic and the 3D human motion estimation problems in Section 4.

1.1. *Related work*

The problem of dynamical state prediction frequently arises in many application domains including computer vision and pattern recognition. The human motion estimation is one of the most interesting tasks, where the dynamic model based approaches including [8,20,2] have opened a promising direction. Beyond the generative modeling and learning, the discriminative approaches have emerged recently exhibiting outstanding prediction accuracy in a variety of situations. In robotics community, [21] empirically studied several objectives for learning of dynamical systems. In contrast to [21]'s ad hoc optimization method, [1] provided efficient gradient-based optimization algorithms for discriminative objectives to dynamical systems. In parallel, several different discriminative models have been successfully applied to dynamic pose estimation problems under diverse circumstances [9–11].

Some of the recent approaches aimed at modeling the latent intrinsic structure of the state variables using nonlinear models such as the Gaussian process latent variable model [22] or its extensions by incorporating latent dynamics or spatial/temporal constraints [23,24]. Although these models can faithfully represent the complex high-order, nonlinear nature of the real motion dynamics, one of their main drawbacks is that one needs to solve difficult non-convex optimization problems with respect to a large number of parameters (e.g., latent training points, kernel, and model parameters), where the overall generalization performance is often highly sensitive to the choice of these parameters. Thus, we focus on the linear dynamic models that can yield high prediction accuracy and favorable tractability if learned properly. The main advantage of linear dynamic models against nonlinear ones is that the linear models have smaller numbers of model parameters which can often yield better generalization performance given a restricted number of data samples. Moreover, the inference or decoding for the linear models can be much faster and exact, whilst the nonlinear models often suffer from computationally intensive particle sampling or nonlinear optimization.

However, most approaches assume the fully supervised settings, requiring costly state (label) acquisition steps and unable to exploit cheap abundant unlabeled data. Although the recent work of [25] tackled the semi-supervised pose estimation problem similar to ours, the major difference is that they deal with static regression, i.e., non-dynamic data, predicting a pose from a still image. Furthermore, whereas they focus on the low-level representation and feature extraction of robust and discriminative image descriptors with emphasis on achieving resistance to background clutter, we aim at developing semi-supervised learning algorithms in general for the challenging dynamic output settings. In this paper, we suggest discriminative semi-supervised algorithms for dynamical systems that can be framed in the unifying semi-supervised regularization framework by incorporating discriminative objectives for both labeled and unlabeled data. One of our approaches, the self-training, is popular for the static (non-dynamic) classification [26–28], while we apply it to dynamical systems, and provide a unifying view of coordinate ascent optimization for a proper discriminative objective. The entropy minimization is motivated from [18]'s semi-supervised learning of CRFs for structured output classification in *discrete* state domains. We extend it to the real-valued multivariate state domain, deriving the nested recursive inference algorithm for the Gaussian random fields defined by the LDS model.

2. Backgrounds: linear dynamical systems

The linear dynamical system (LDS) is a generative sequence model $P(\mathbf{X},\mathbf{Y})$ that places a 1st-order linear Gaussian dynamics on the states $\mathbf{Y}$ and the linear Gaussian emission for the measurement $\mathbf{x}_t$ given $\mathbf{y}_t$. The conditional densities of LDS are defined as

$$\mathbf{y}_1 \sim \mathcal{N}(\mathbf{y}_1;\mathbf{m}_0,\mathbf{V}_0), \quad \mathbf{y}_t|\mathbf{y}_{t-1} \sim \mathcal{N}(\mathbf{y}_t;\mathbf{A}\mathbf{y}_{t-1},\mathbf{\Gamma}), \quad \mathbf{x}_t|\mathbf{y}_t \sim \mathcal{N}(\mathbf{x}_t;\mathbf{C}\mathbf{y}_t,\mathbf{\Sigma}). \tag{2}$$

Here $\theta := \{\mathbf{m}_0,\mathbf{V}_0,\mathbf{A},\mathbf{\Gamma},\mathbf{C},\mathbf{\Sigma}\}$ is the LDS parameter set, where $\mathbf{m}_0 \in \mathbb{R}^{d\times 1}$, $\mathbf{V}_0,\mathbf{A},\mathbf{\Gamma} \in \mathbb{R}^{d\times d}$, $\mathbf{C} \in \mathbb{R}^{p\times d}$, and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$. The joint log-likelihood, $JLL = \log P(\mathbf{X},\mathbf{Y})$ is (up to a constant):

$$\begin{aligned} JLL = -\frac{1}{2}[&(\mathbf{y}_1-\mathbf{m}_0)'\mathbf{V}_0^{-1}(\mathbf{y}_1-\mathbf{m}_0)+\log|\mathbf{V}_0| \\ &+\sum_{t=2}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{y}_{t-1})'\mathbf{\Gamma}^{-1}(\mathbf{y}_t-\mathbf{A}\mathbf{y}_{t-1})+\log|\mathbf{\Gamma}|^{T-1} \\ &+\sum_{t=1}^{T}(\mathbf{x}_t-\mathbf{C}\mathbf{y}_t)'\mathbf{\Sigma}^{-1}(\mathbf{x}_t-\mathbf{C}\mathbf{y}_t)+\log|\mathbf{\Sigma}|^{T}], \end{aligned} \tag{3}$$

where $\mathbf{M}'$ indicates the transpose of the matrix $\mathbf{M}$.

The task of inference is to compute the filtered state densities, $P(\mathbf{y}_t|\mathbf{x}_1,\ldots,\mathbf{x}_t)$ and the smoothed densities, $P(\mathbf{y}_t|\mathbf{X})$. The linear Gaussian assumption of the LDS implies that these filtered/smoothed densities are Gaussian, and can be evaluated in linear time using the well-known Kalman filtering and smoothing algorithms [29]. We denote their means and covariances by $\hat{\mathbf{m}}_t := \mathbb{E}[\mathbf{y}_t|\mathbf{x}_1,\ldots,\mathbf{x}_t]$, $\hat{\mathbf{V}}_t := \mathbb{V}(\mathbf{y}_t|\mathbf{x}_1,\ldots,\mathbf{x}_t)$, $\mathbf{m}_t := \mathbb{E}[\mathbf{y}_t|\mathbf{X}]$, $\mathbf{V}_t := \mathbb{V}(\mathbf{y}_t|\mathbf{X})$, and $\mathbf{\Sigma}_{t,t-1} :=$ Cov $(\mathbf{y}_t,\mathbf{y}_{t-1}|\mathbf{X})$.

The task of learning an LDS is to find θ that maximizes a desired objective function. In the supervised setting, for the given training data $\mathcal{L} = \{(\mathbf{X}^i = \mathbf{x}_1^i \cdots \mathbf{x}_{T_i}^i,\mathbf{Y}^i = \mathbf{y}_1^i \cdots \mathbf{y}_{T_i}^i)\}_{i=1}^n$, the standard maximum likelihood (ML) learning maximizes the joint log-likelihood, $\sum_{i=1}^n JLL(\mathbf{X}^i,\mathbf{Y}^i)$. The ML learning has a closed-form solution by setting the gradients of (3) to 0. For instance, $\mathbf{C}^* = [\sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{x}_t^i\ \mathbf{x}_t^{i\prime}] \cdot [\sum_{i=1}^n \sum_{t=1}^{T_i} \mathbf{y}_t^i\ \mathbf{y}_t^{i\prime}]^{-1}$, where T_i is the length of the i-th sequence. The gradients of the joint log-likelihood are derived as follows:

$$\begin{aligned} &\frac{\partial JLL}{\partial \mathbf{m}_0} = \mathbf{V}_0^{-1}(\mathbf{y}_1-\mathbf{m}_0), \quad \frac{\partial JLL}{\partial \mathbf{V}_0^{-1}} = \frac{1}{2}\mathbf{V}_0-\frac{1}{2}(\mathbf{y}_1-\mathbf{m}_0)(\mathbf{y}_1-\mathbf{m}_0)', \\ &\frac{\partial JLL}{\partial \mathbf{A}} = \mathbf{\Gamma}^{-1}\cdot\sum_{t=2}^{T}(\mathbf{y}_t\mathbf{y}'_{t-1}-\mathbf{A}\mathbf{y}_{t-1}\mathbf{y}'_{t-1}), \\ &\frac{\partial JLL}{\partial \mathbf{\Gamma}^{-1}} = \frac{T-1}{2}\mathbf{\Gamma}-\frac{1}{2}\sum_{t=2}^{T}(\mathbf{y}_t-\mathbf{A}\mathbf{y}_{t-1})(\mathbf{y}_t-\mathbf{A}\mathbf{y}_{t-1})', \\ &\frac{\partial JLL}{\partial \mathbf{C}} = \mathbf{\Sigma}^{-1}\cdot\sum_{t=1}^{T}(\mathbf{x}_t\mathbf{y}'_t-\mathbf{C}\mathbf{y}_t\mathbf{y}'_t), \quad \frac{\partial JLL}{\partial \mathbf{\Sigma}^{-1}} = \frac{T}{2}\mathbf{\Sigma}-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t-\mathbf{C}\mathbf{y}_t)(\mathbf{y}_t-\mathbf{C}\mathbf{y}_t)'. \end{aligned} \tag{4}$$

The ML learning of the generative model is intended to fit the model to data jointly on $\mathbf{X}$ and $\mathbf{Y}$. However, in the motion estimation task, we are more interested in finding a model that yields a high accuracy in predicting $\mathbf{Y}$ from $\mathbf{X}$, an objective *not* achieved by the ML learning in general. In our earlier work [1,12], we suggested two discriminative learning algorithms that explicitly focus on the desired goal of prediction.

2.1. Conditional maximum likelihood (CML)

The CML learning maximizes the conditional likelihood of the entire state sequence $\mathbf{Y}$ given the measurement sequence $\mathbf{X}$. We locally optimize the non-convex objective $CLL = \log P(\mathbf{Y}|\mathbf{X})$ using gradient search:

$$\frac{\partial CLL}{\partial \theta} = \frac{\partial \log P(\mathbf{X},\mathbf{Y})}{\partial \theta} - \mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}\left[\frac{\partial \log P(\mathbf{X},\mathbf{Y})}{\partial \theta}\right]. \tag{5}$$

Note that the gradient is the difference between the derivative of the joint log-likelihood (also known as the Fisher score) evaluated on the complete data $(\mathbf{X},\mathbf{Y})$ and the expected one by the current model on the observation $\mathbf{X}$. The Fisher score can be easily computed from (4), while the expectation requires the state inference $P(\mathbf{Y}|\mathbf{X})$ which can be done by the Kalman smoothing algorithm.

2.2. Slicewise conditional maximum likelihood (SCML)

The goal of the CML learning was to find a model that minimizes the state estimation error at the *entire* sequence level. However, it is more natural to consider the prediction error at each *time slice* individually, namely, maximizing the slicewise conditional log-likelihood, $SCLL = (1/T)\sum_{t=1}^T \log P(\mathbf{y}_t|\mathbf{X})$. This slicewise CML (or SCML) learning, directly related to the minimization of the Hamming distance, was previously proposed as an alternative cost function for CRFs in discrete state domains [30]. The SCML learning of the LDS is shown to perform well especially when the model structure assumption is incorrect [1]. However, its two-pass forward/backward inference algorithm ($O(T^2)$ time) may be computationally demanding for long sequences. More details about the optimization are described in [1,12].

3. Discriminative semi-supervised learning

In this section, we propose new discriminative semi-supervised learning algorithms that can exploit the unlabeled data $\mathcal{U} = \{\mathbf{X}^j\}_{j=n+1}^m$ in conjunction with the labeled data $\mathcal{L} = \{(\mathbf{X}^i,\mathbf{Y}^i)\}_{i=1}^n$. As discussed in the introduction, we follow the general semi-supervised learning framework in Eq. (1), where the learning objective to be maximized over the LDS parameters θ is comprised of the supervised criterion $O(\theta,\mathcal{L})$ and the negatively scaled regularization term $R(\theta,\mathcal{U})$ defined over the unlabeled data. For the labeled part, we let the supervised discriminative learning criteria (i.e., either CML or SCML) take the place of $O(\theta,\mathcal{L})$. In the following, we focus on the regularization term $R(\theta,\mathcal{U})$ on the unlabeled part by suggesting two approaches: self-training and entropy minimization.

3.1. Self-training

A fairly standard way to incorporate unlabeled data in the generative probabilistic model would be to treat the unlabeled data as *missing* data and to maximize the marginal log-likelihood. This is exactly equivalent to letting $R(\theta,\mathcal{U}) = -\sum_{j\in\mathcal{U}} \log P(\mathbf{X}^j;\theta)$. The maximization of the marginal log-likelihood can be typically done by the well-known expectation maximization (EM) algorithm [31]. Despite its popularity, the integration over the unlabeled (latent) states $\mathbf{Y}$ is purely generative, merely aiming at fitting of the model to overall data. To have a more discriminative model, one can regard the LDS model as a state predictor $P(\mathbf{Y}|\mathbf{X})$, and incorporate it into a semi-supervised learning framework.

We adopt the self-training strategy [26–28], one of the most popular semi-supervised approaches that can make use of the unlabeled data in a sensible manner. In the pattern recognition literature, it has been similarly studied named as decision-directed algorithm [32]. The self-training algorithm for the LDS model works as follows:

1. Initially, we let $\mathcal{D}=\mathcal{L}$.
2. Repeat the following steps until convergence:
   (a) Do supervised LDS learning with $\mathcal{D}$ to have a new model θ^{new}, i.e., $\theta^{new}=\text{argmax}_\theta O(\theta,\mathcal{D})$.
   (b) Determine the labels for $\mathcal{U}$ using the current model, i.e., $\hat{\mathbf{Y}}^j=\text{argmax}_\mathbf{Y} P(\mathbf{Y}|\mathbf{X}^j;\theta^{new})$ for $j\in\mathcal{U}$. The maximization can be done by the Kalman smoothing.
   (c) Add $\{(\mathbf{X}^j,\hat{\mathbf{Y}}^j)\}_{j\in\mathcal{U}}$ to $\mathcal{D}$. (Or, replace the old ones.)

Although the self-training seems to have little connection to the regularization framework, we will show that it indeed belongs to the framework, derived as coordinate ascent optimization of a proper discriminative objective over both model parameters and the unlabeled state sequences. First, when we employ the CML objective for $O(\theta,\mathcal{L})$, we consider the following optimization problem:

$$\max_{\theta,\{\mathbf{Y}^j\}}\sum_{i\in\mathcal{L}}\log P(\mathbf{Y}^i|\mathbf{X}^i)+\lambda\cdot\sum_{j\in\mathcal{U}}\log P(\mathbf{Y}^j|\mathbf{X}^j). \tag{6}$$

Note that we use the same objective form (CML) for the unlabeled part $R(\theta,\mathcal{U})$, where the unknown labels $\{\mathbf{Y}^j\}$ are also optimized simultaneously with the model parameters θ. In the coordinate ascent manner, when we fix θ, the optimization over $\mathbf{Y}^j$ is confined to the second term, and it corresponds to Step-2(b) in the self-training. Also, when we fix $\mathbf{Y}^j$, the optimization of (6) over θ is exactly the same as Step-2(a) (assuming λ set to 1). In summary, the self-training with the CML learning criterion is equivalent to the coordinate ascent optimization of (6) with $\lambda=1$ and the regularization term defined as

$$R(\theta,\mathcal{U})=-\sum_{j\in\mathcal{U}}\max_{\mathbf{Y}^j}\log P(\mathbf{Y}^j|\mathbf{X}^j). \tag{7}$$

Similarly, for the SCML criterion, the self-training can be equivalently achieved by letting[1]:

$$R(\theta,\mathcal{U})=-\sum_{j\in\mathcal{U}}\frac{1}{T_j}\sum_{t=1}^{T_j}\max_{\mathbf{y}_t^j}\log P(\mathbf{y}_t^j|\mathbf{X}^j). \tag{8}$$

Hence, the self-training can be seen as a special case ($\lambda=1$) of the general semi-supervised regularization framework, which admits a discriminative objective identical for both labeled and unlabeled data.

3.2. Entropy minimization

Among many existing semi-supervised learning methods, one of the most successful approaches is the entropy minimization whose framework was originally proposed by [17] for the static classification problem. Recently it was applied to the structured output classification problem [18,19], where they minimize the entropy of the conditional distribution of the discrete label sequence for the unlabeled data. In this section, we further extend it to the LDS models, that is, the dynamic continuous state regression setting of the Gaussian random fields.

The main idea is to employ the entropy of the conditional density for the unlabeled data as the regularization term, i.e., $R(\theta,\mathcal{U})=\sum_{j\in\mathcal{U}}-\int_\mathbf{Y}P(\mathbf{Y}|\mathbf{X}^j)\log P(\mathbf{Y}|\mathbf{X}^j)$, thus forcing the model to have minimal uncertainty in predicting labels for the unlabeled data. Its theoretical basis is motivated from the minimization of the Kullback–Leibler divergence between the model-induced conditional density and the empirical conditional density on the unlabeled data, which in the classification cases, is shown to guide decision boundaries to lie on low-density regions [15]. In the structured output *regression* setting, the notion can be similarly extended.

Having employed the negative entropy regularization term, the semi-supervised LDS learning can be formulated as the following optimization problem:

$$\max_\theta O(\theta,\mathcal{L})+\lambda\cdot\sum_{j\in\mathcal{U}}\int_\mathbf{Y}P(\mathbf{Y}|\mathbf{X}^j)\log P(\mathbf{Y}|\mathbf{X}^j)\,d\mathbf{Y}, \tag{9}$$

where we take either CML or SCML objective described in Sections 2.1 and 2.2 for the labeled objective $O(\theta,\mathcal{L})$. Not only the CML/SCML objectives, but the entropy regularization term is also known to be non-convex in terms of θ, meaning that (9) has many local maxima. However, we will show that gradient search can significantly improve the prediction performance of the model over the supervised solution that simply ignores the regularization term. The optimization of CML/SCML has been studied in our earlier work, and the rest of the section is devoted for describing how to optimize the regularization part (i.e., evaluating the negative entropy and its gradient wrt θ).

First, evaluating a negative entropy is straightforward when the Kalman smoothing is done on the unlabeled data. That is,

$$\begin{aligned}\int_\mathbf{Y}P(\mathbf{Y}|\mathbf{X})\log P(\mathbf{Y}|\mathbf{X})\,d\mathbf{Y}&=\mathbb{E}[\log P(\mathbf{X},\mathbf{Y})-\log P(\mathbf{X})]\\&=\mathbb{E}[\log P(\mathbf{X},\mathbf{Y})]-\log P(\mathbf{X}).\end{aligned} \tag{10}$$

The expectations above (and, in fact, all other expectations in this section) are taken over $P(\mathbf{Y}|\mathbf{X})$. The measurement likelihood in (10), i.e., $\log P(\mathbf{X})$, can be readily available from the Kalman filtering procedure, while the expectation of the joint log-likelihood (*JLL*) can be easily computed since *JLL* is a sum of adjoining 1st/2nd-order moments (i.e., $\mathbf{y}_t\mathbf{y}'_{t-1}$, $\mathbf{y}_t\mathbf{y}'_t$, or $\mathbf{y}_t$) from (3). For instance, $\mathbb{E}[\mathbf{y}_t\mathbf{y}'_{t-1}]=\mathbf{\Sigma}_{t,t-1}+\mathbf{m}_t\mathbf{m}'_{t-1}$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}'_t]=\mathbf{V}_t+\mathbf{m}_t\mathbf{m}'_t$, and $\mathbb{E}[\mathbf{y}_t]=\mathbf{m}_t$.

Computing the gradient of the negative entropy is rather more complicated. From the following identity (see Appendix A for the detailed derivation),

$$\begin{aligned}\frac{\partial}{\partial\theta}\int_\mathbf{Y}P(\mathbf{Y}|\mathbf{X})\log P(\mathbf{Y}|\mathbf{X})\,d\mathbf{Y}&=\mathbb{E}\left[\log P(\mathbf{X},\mathbf{Y})\cdot\frac{\partial}{\partial\theta}\log P(\mathbf{X},\mathbf{Y})\right]\\&\quad-\mathbb{E}[\log P(\mathbf{X},\mathbf{Y})]\cdot\mathbb{E}\left[\frac{\partial}{\partial\theta}\log P(\mathbf{X},\mathbf{Y})\right]\end{aligned} \tag{11}$$

we see that the main complexity lies in computing the first term, $\mathbb{E}[\log P(\mathbf{X},\mathbf{Y})\cdot(\partial/\partial\theta)\log P(\mathbf{X},\mathbf{Y})]$, since the second term is easy to compute once one has the state posteriors from the Kalman smoothing.[2] We then discuss how to compute the first term in greater detail. From the adjoining 1st/2nd-order moment forms of $\log P(\mathbf{X},\mathbf{Y})$ and $(\partial/\partial\theta)\log P(\mathbf{X},\mathbf{Y})$ from (3) and (4), respectively, it is not difficult to see that the expectation is a sum of 4th-order moments composed of $(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1})$ for $1\le s<t\le T$. Thus one has to evaluate the density $P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$, and this can be done in a dynamic programming fashion: for each $s=1,\dots,T-1$, we compute

[1] Since $P(\mathbf{Y}|\mathbf{X})$ is jointly Gaussian over $\mathbf{y}_1,\dots,\mathbf{y}_T$, and the maximizer of the density coincides with its mean, the Kalman smoothing results are also the solution to the maximization problem in (8).

[2] That is, $\mathbb{E}[(\partial/\partial\theta)\log P(\mathbf{X},\mathbf{Y})]$ is a sum of adjoining 1st-/2nd-order moment expectations from (4), while $\mathbb{E}[\log P(\mathbf{X},\mathbf{Y})]$ is exactly the negative entropy.

$P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$ recursively on t (starting from $t=s+1$ until $t=T$). Note that the overall procedure requires $O(T^2)$ time as we eventually fill in a $(T\times T)$ table for each entry (s,t).

Our derivation for $P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$ is based on the general message passing in undirected graphical models [3,33], where the graph structure of the LDS model conforms to the 1st-order Markov chain. For this purpose, we first reparameterize the LDS model in an equivalent log-linear form in the next section, which facilitates deriving the message passing. Interestingly, the standard forward/backward message passing in this log-linear model can be seen as an alternative inference that can replace the traditional Kalman smoothing. In Section 3.2.2, we describe how the desired density $P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$ can be obtained from the nested recursion on the log-linear model.

3.2.1. *Log-linear representation of LDS model*

We form an LDS-equivalent log-linear conditional model $P(\mathbf{Y}|\mathbf{X})$, where we regard $\mathbf{X}$ as constant since the input measurement sequence is always given. As $P(\mathbf{Y}|\mathbf{X})\propto P(\mathbf{X},\mathbf{Y})$, we rewrite the joint log-likelihood of the LDS as a linear form by introducing new parameters. More specifically, we introduce new parameters $\{(\mathbf{v}_t\in\mathbb{R}^{d\times 1},\mathbf{Q}_t\in\mathbb{R}^{d\times d},\mathbf{S}_t\in\mathbb{R}^{d\times d})\}_{t=1}^T$ such that the joint log-likelihood in (3) can be expressed as a linear function of the new parameters,

$$\log P(\mathbf{X},\mathbf{Y})=\sum_{t=1}^T \mathbf{v}'_t\mathbf{y}_t-\sum_{t=2}^T \mathbf{y}'_t\mathbf{Q}_t\mathbf{y}_{t-1}-\frac{1}{2}\sum_{t=1}^T \mathbf{y}'_t\mathbf{S}_t\mathbf{y}_t+\rho(\theta,\mathbf{X}). \tag{12}$$

Then there exists the following correspondence between the LDS parameters and the new parameters:

$$\mathbf{v}_1=\mathbf{V}_0^{-1}\mathbf{m}_0+\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{x}_1,\quad \mathbf{v}_t=\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{x}_t\quad(\text{for } t\geq 2),$$
$$\mathbf{Q}_t=-\mathbf{\Gamma}^{-1}\mathbf{A}\quad(\text{for }\ 1\leq t\leq T)$$

$$\mathbf{S}_1=\mathbf{V}_0^{-1}+\mathbf{A}'\mathbf{\Gamma}^{-1}\mathbf{A}+\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C},$$
$$\mathbf{S}_t=\mathbf{\Gamma}^{-1}+\mathbf{A}'\mathbf{\Gamma}^{-1}\mathbf{A}+\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}\quad(\text{for }\ 2\leq t\leq T-1)$$

$$\mathbf{S}_T=\mathbf{\Gamma}^{-1}+\mathbf{C}'\mathbf{\Sigma}^{-1}\mathbf{C}, \tag{13}$$

where $\mathbf{Y}$-independent $\rho(\theta,\mathbf{X})=-\frac{1}{2}(\log|\mathbf{V}_0|+(T-1)\log|\mathbf{\Gamma}|+T\log|\mathbf{\Sigma}|+\mathbf{m}'_0\mathbf{V}_0^{-1}\mathbf{m}_0+\sum_{t=1}^T\mathbf{x}'_t\mathbf{\Sigma}^{-1}\mathbf{x}_t)$.

The parameterization in (12) defines the log-linear model,

$$P(\mathbf{Y}|\mathbf{X})\propto\exp\left(\sum_{t=1}^T \mathbf{v}'_t\mathbf{y}_t-\sum_{t=2}^T \mathbf{y}'_t\mathbf{Q}_t\mathbf{y}_{t-1}-\frac{1}{2}\sum_{t=1}^T \mathbf{y}'_t\mathbf{S}_t\mathbf{y}_t\right) \tag{14}$$

from which the inference $P(\mathbf{y}_t|\mathbf{X})$ and $P(\mathbf{y}_t,\mathbf{y}_{t-1}|\mathbf{X})$, traditionally computed by the Kalman smoothing, can now be obtained by running the standard message passing (also known as the forward/backward recursion). More specifically, for the given $\mathbf{X}$, the potential function $M_t(\cdot)$ is defined on the adjoining state pair $(\mathbf{y}_t,\mathbf{y}_{t-1})$ (i.e., the *clique* in the chain structured graphical model) at time $t\geq 2$ as

$$M_t(\mathbf{y}_t,\mathbf{y}_{t-1})=\exp(-\tfrac{1}{2}\mathbf{y}'_t\mathbf{S}_t\mathbf{y}_t-\mathbf{y}'_t\mathbf{Q}_t\mathbf{y}_{t-1}+\mathbf{v}'_t\mathbf{y}_t), \tag{15}$$

with $M_1(\mathbf{y}_1)=\exp(-\frac{1}{2}\mathbf{y}'_1\mathbf{S}_1\mathbf{y}_1+\mathbf{v}'_1\mathbf{y}_1)$ initially. We then recursively define the forward messages with the initial condition, $\alpha_1(\mathbf{y}_1)=M_1(\mathbf{y}_1)$, and for $t=2,\ldots,T$,

$$\alpha_t(\mathbf{y}_t)=\int_{\mathbf{y}_{t-1}}\alpha_{t-1}(\mathbf{y}_{t-1})\cdot M_t(\mathbf{y}_t,\mathbf{y}_{t-1})\,d\mathbf{y}_{t-1}. \tag{16}$$

Since $\alpha_t(\mathbf{y}_t)$ is an unnormalized Gaussian, it can be represented as a triplet, $(r_t,\mathbf{P}_t,\mathbf{q}_t)\in(\mathbb{R},\mathbb{R}^{d\times d},\mathbb{R}^d)$, implying that $\alpha_t(\mathbf{y}_t)=r_t\exp(-\frac{1}{2}\mathbf{y}'_t\mathbf{P}_t\mathbf{y}_t+\mathbf{q}'_t\mathbf{y}_t)$. Following the recursion in (16), we can compute $(r_t,\mathbf{P}_t,\mathbf{q}_t)$ recursively as follows:

$$t=1;\quad r_1=1,\quad \mathbf{P}_1=\mathbf{S}_1,\quad \mathbf{q}_1=\mathbf{v}_1,$$

$$t\geq 2;\quad r_t=r_{t-1}\cdot|2\pi\mathbf{P}_{t-1}^{-1}|^{1/2}\cdot e^{(1/2)\mathbf{q}'_{t-1}\mathbf{P}_{t-1}^{-1}\mathbf{q}_{t-1}},$$

$$\mathbf{P}_t=\mathbf{S}_t-\mathbf{Q}_t\mathbf{P}_{t-1}^{-1}\mathbf{Q}'_t,\quad \mathbf{q}_t=\mathbf{v}_t-\mathbf{Q}_t\mathbf{P}_{t-1}^{-1}\mathbf{q}_{t-1}. \tag{17}$$

It is important to notice that (17) holds since the integrations in (16) are well defined (not infinity) due to the joint Gaussianity of $P(\mathbf{Y}|\mathbf{X})$.

Similarly, we can define the backward messages with the initial, $\beta_T(\mathbf{y}_T)=1$, and for $t<T$,

$$\beta_t(\mathbf{y}_t)=\int_{\mathbf{y}_{t+1}}M_{t+1}(\mathbf{y}_{t+1},\mathbf{y}_t)\cdot\beta_{t+1}(\mathbf{y}_{t+1})d\mathbf{y}_{t+1}. \tag{18}$$

The triplet, $(h_t,\mathbf{F}_t,\mathbf{g}_t)\in(\mathbb{R},\mathbb{R}^{d\times d},\mathbb{R}^d)$, for representing $\beta_t(\mathbf{y}_t)=h_t\exp(-\frac{1}{2}\mathbf{y}'_t\mathbf{F}_t\mathbf{y}_t+\mathbf{g}'_t\mathbf{y}_t)$, can be computed as

$$t=T;\quad h_T=1,\quad \mathbf{F}_T=0_{d\times d},\quad \mathbf{g}_T=0_{d\times 1},$$

$$t<T;\quad h_t=h_{t+1}\cdot|2\pi\tilde{\mathbf{F}}_{t+1}^{-1}|^{1/2}\cdot e^{(1/2)\tilde{\mathbf{g}}'_{t+1}\tilde{\mathbf{F}}_{t+1}^{-1}\tilde{\mathbf{g}}_{t+1}},$$

$$\mathbf{F}_t=-\mathbf{Q}'_{t+1}\tilde{\mathbf{F}}_{t+1}^{-1}\mathbf{Q}_{t+1},\quad \mathbf{g}_t=-\mathbf{Q}'_{t+1}\tilde{\mathbf{F}}_{t+1}^{-1}\tilde{\mathbf{g}}_{t+1}, \tag{19}$$

where $\tilde{\mathbf{F}}_t:=\mathbf{F}_t+\mathbf{S}_t$ and $\tilde{\mathbf{g}}_t:=\mathbf{g}_t+\mathbf{v}_t$. Likewise, the integrations in (18) are well defined.

From the forward/backward messages, the inference can be obtained using the fact that $P(\mathbf{y}_t|\mathbf{X})\propto\alpha_t(\mathbf{y}_t)\cdot\beta_t(\mathbf{y}_t)$ and $P(\mathbf{y}_t,\mathbf{y}_{t-1}|\mathbf{X})\propto\alpha_{t-1}(\mathbf{y}_{t-1})\cdot M_t(\mathbf{y}_t,\mathbf{y}_{t-1})\cdot\beta_t(\mathbf{y}_t)$ as follows:

$$\mathbf{m}_t=\mathbb{E}[\mathbf{y}_t]=(\mathbf{P}_t+\mathbf{F}_t)^{-1}\cdot(\mathbf{q}_t+\mathbf{g}_t),$$

$$\mathbf{V}_t=\mathbb{V}(\mathbf{y}_t)=(\mathbf{P}_t+\mathbf{F}_t)^{-1},$$

$$\mathbf{\Sigma}_{t,t-1}=\mathrm{Cov}(\mathbf{y}_t,\mathbf{y}_{t-1})=-\tilde{\mathbf{F}}_t^{-1}\mathbf{Q}_t(\mathbf{P}_{t-1}-\mathbf{Q}'_t\tilde{\mathbf{F}}_t^{-1}\mathbf{Q}_t)^{-1},$$

which are equivalent to those computed by Kalman smoothing.[3]

3.2.2. *Nested recursion for $P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$*

We compute the desired density $P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})$ for $1\leq s<t\leq T$ from the joint density of (14) using the following marginalization:

$$\begin{aligned}P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X})&=\int_{\mathbf{Y}\backslash\{\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}\}}P(\mathbf{Y}|\mathbf{X})\,d\mathbf{Y}\\&\propto\int_{\mathbf{Y}\backslash\{\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}\}}e^{\sum_{t=1}^T\mathbf{v}'_t\mathbf{y}_t-\sum_{t=2}^T\mathbf{y}'_t\mathbf{Q}_t\mathbf{y}_{t-1}-(1/2)\sum_{t=1}^T\mathbf{y}'_t\mathbf{S}_t\mathbf{y}_t},\end{aligned} \tag{20}$$

where $A\backslash B$ indicates *set-minus* that excludes the set B from A. The integration in (20) can be decomposed into five parts as shown in the following diagram:

$$\underbrace{\mathbf{y}_1\sim\mathbf{y}_{s-2}}_{=\alpha_{s-1}(\mathbf{y}_{s-1})}|\ \underbrace{\mathbf{y}_{s-1},\mathbf{y}_s}_{=M_s(\mathbf{y}_s,\mathbf{y}_{s-1})}\ |\underbrace{\mathbf{y}_{s+1}\sim\mathbf{y}_{t-2}}_{=N_{t-1,s}(\mathbf{y}_{t-1},\mathbf{y}_s)}|\ \underbrace{\mathbf{y}_{t-1},\mathbf{y}_t}_{=M_t(\mathbf{y}_t,\mathbf{y}_{t-1})}\ |\underbrace{\mathbf{y}_{t+1}\sim\mathbf{y}_T}_{=\beta_t(\mathbf{y}_t)}.$$

The first group indicates marginalization over $\mathbf{y}_1\sim\mathbf{y}_{s-2}$, which is exactly the same to the forward message $\alpha_{s-1}(\mathbf{y}_{s-1})$ by construction. Similarly the last group that marginalizes over $\mathbf{y}_{t+1}\sim\mathbf{y}_T$ is equal to $\beta_t(\mathbf{y}_t)$. The second and the fourth groups involve the exponential terms dependent on $(\mathbf{y}_{s-1},\mathbf{y}_s)$ and $(\mathbf{y}_{t-1},\mathbf{y}_t)$, respectively, which can be further reduced to $M_s(\mathbf{y}_s,\mathbf{y}_{s-1})$ and $M_t(\mathbf{y}_t,\mathbf{y}_{t-1})$ by definition. The remaining part is the marginalization over the variables $\mathbf{y}_{s+1}\sim\mathbf{y}_{t-2}$ in the middle, which we define as

$$N_{t-1,s}(\mathbf{y}_{t-1},\mathbf{y}_s)=\int_{\mathbf{y}_{s+1}\cdots\mathbf{y}_{t-2}}\exp\left(\sum_{l=s+1}^{t-1}\left(-\frac{1}{2}\mathbf{y}'_l\mathbf{S}_l\mathbf{y}_l-\mathbf{y}'_l\mathbf{Q}_l\mathbf{y}_{l-1}+\mathbf{v}'_l\mathbf{y}_l\right)\right) \tag{21}$$

[3] Interestingly, one benefit of the above inference is the reduced time complexity: the message passing takes $O(T\cdot(d^3+dp))$ time, while the Kalman smoothing requires $O(T\cdot(d^3+p^3+dp))$. The linear time in the measurement dimension (p) allows us to incorporate high-dimensional, possibly nonlinearly expanded, measurement features.

for $s=1,\ldots,T-1$ and $t=s+1,\ldots,T$. Using this definition, we can express the posterior density as

$$P(\mathbf{y}_t,\mathbf{y}_{t-1},\mathbf{y}_s,\mathbf{y}_{s-1}|\mathbf{X}) \propto \alpha_{s-1}(\mathbf{y}_{s-1})\cdot M_s(\mathbf{y}_s,\mathbf{y}_{s-1})\cdot N_{t-1,s}(\mathbf{y}_{t-1},\mathbf{y}_s) \cdot M_t(\mathbf{y}_t,\mathbf{y}_{t-1})\cdot \beta_t(\mathbf{y}_t). \tag{22}$$

We describe how to compute $N_{t,s}(\mathbf{y}_t,\mathbf{y}_s)$ (note that we plug t in the place of $t-1$ in (21)). We do this via nested recursion, namely recursion over $t=s,\ldots,T-1$ for each $s=1,\ldots,T-1$. First, since $N_{t,s}(\mathbf{y}_t,\mathbf{y}_s)$ is an unnormalized Gaussian, we represent it as 6-tuple, $(k_{t,s},\mathbf{B}_{t,s},\mathbf{E}_{t,s},\mathbf{H}_{t,s},\mathbf{u}_{t,s},\mathbf{w}_{t,s}) \in (\mathbb{R},\mathbb{R}^{d\times d},\mathbb{R}^{d\times d},\mathbb{R}^{d\times d},\mathbb{R}^{d\times 1},\mathbb{R}^{d\times 1})$, implying that

$$N_{t,s}(\mathbf{y}_t,\mathbf{y}_s) = k_{t,s}\exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{y}_t\\ \mathbf{y}_s\end{pmatrix}'\begin{pmatrix}\mathbf{B}_{t,s} & \mathbf{E}_{t,s}\\ \mathbf{E}'_{t,s} & \mathbf{H}_{t,s}\end{pmatrix}\begin{pmatrix}\mathbf{y}_t\\ \mathbf{y}_s\end{pmatrix}+\begin{pmatrix}\mathbf{u}_{t,s}\\ \mathbf{w}_{t,s}\end{pmatrix}'\begin{pmatrix}\mathbf{y}_t\\ \mathbf{y}_s\end{pmatrix}\right). \tag{23}$$

Then, it is not difficult to see that it can be computed by the following recursion for $t=s+2,\ldots,T-1$:

$$\begin{aligned}
k_{t,s} &= k_{t-1,s}\cdot|2\pi\mathbf{B}^{-1}_{t-1,s}|^{1/2}\cdot\exp\left(\frac{1}{2}\mathbf{u}'_{t-1,s}\mathbf{B}^{-1}_{t-1,s}\mathbf{u}_{t-1,s}\right),\\
\mathbf{B}_{t,s} &= \mathbf{S}_t-\mathbf{Q}_t\mathbf{B}^{-1}_{t-1,s}\mathbf{Q}'_t,\\
\mathbf{E}_{t,s} &= -\mathbf{Q}_t\mathbf{B}^{-1}_{t-1,s}\mathbf{E}_{t-1,s},\\
\mathbf{H}_{t,s} &= \mathbf{H}_{t-1,s}-\mathbf{E}'_{t-1,s}\mathbf{B}^{-1}_{t-1,s}\mathbf{E}_{t-1,s},\\
\mathbf{u}_{t,s} &= \mathbf{v}_t-\mathbf{Q}_t\mathbf{B}^{-1}_{t-1,s}\mathbf{u}_{t-1,s},\\
\mathbf{w}_{t,s} &= \mathbf{w}_{t-1,s}-\mathbf{E}'_{t-1,s}\mathbf{B}^{-1}_{t-1,s}\mathbf{u}_{t-1,s},
\end{aligned} \tag{24}$$

with the initial conditions:

$$\begin{aligned}
&(t=s)\quad \mathbf{B}_{s,s}=\mathbf{E}_{s,s}=\mathbf{H}_{s,s}=0_{d\times d},\\
&\qquad\quad k_{s,s}=1,\quad \mathbf{u}_{s,s}=\mathbf{w}_{s,s}=0_{d\times 1}.\\
&(t=s+1)\quad \mathbf{B}_{s+1,s}=\mathbf{S}_{s+1},\quad \mathbf{E}_{s+1,s}=\mathbf{Q}_{s+1},\quad \mathbf{H}_{s+1,s}=0_{d\times d},\\
&\qquad\qquad k_{s+1,s}=1,\quad \mathbf{u}_{s+1,s}=\mathbf{v}_{s+1},\quad \mathbf{w}_{s+1,s}=0_{d\times 1}.
\end{aligned}$$

Although the posterior density can be explicitly computed from (22) once $N_{t,s}(\mathbf{y}_t,\mathbf{y}_s)$ is evaluated, due to the joint Gaussianity of the density, what we actually require in performing expectation of 4th order moments is the pairwise covariances, in particular, $\mathrm{cov}(\mathbf{y}_\tau,\mathbf{y}_\sigma|\mathbf{X})$, where $\tau=t$ or $t-1$, and $\sigma=s$ or $s-1$. This can be easily derived from (22) by marginalizing out nuisance variables. In a nutshell, the covariance has the following formula:

$$\mathrm{cov}(\mathbf{y}_\tau,\mathbf{y}_\sigma|\mathbf{X}) = -(\mathbf{F}_\tau+\mathbf{B}_{\tau,\sigma})^{-1}\mathbf{E}_{\tau,\sigma}(\mathbf{P}_\sigma+\mathbf{H}_{\tau,\sigma}-\mathbf{E}'_{\tau,\sigma}(\mathbf{F}_\tau+\mathbf{B}_{\tau,\sigma})^{-1}\mathbf{E}_{\tau,\sigma})^{-1}. \tag{25}$$

4. Evaluation

The performance of the proposed discriminative semi-supervised approaches is evaluated on both synthetic data and the human motion estimation dataset. To distinguish different learning algorithms for the LDS, we use the notation "*OBJ-SS*", where *OBJ* represents the objective for the labeled data (i.e., $O(\theta,\mathcal{L})$ in (1)), and *SS* indicates the regularization term $R(\theta,\mathcal{U})$ on the unlabeled data. Specifically, we use the following acronyms. First, *OBJ* takes either of:

- *ML*, traditional maximum likelihood learning.
- *CML*, conditional maximum likelihood of Section 2.1.
- *SCML*, slicewise CML of Section 2.2.

And, the semi-supervised criterion *SS* is chosen from:

- *Sup*, supervised learning that simply ignores the unlabeled data, i.e., $R(\theta,\mathcal{U})=0$.
- *MLM*, Standard generative approach that maximizes the marginal likelihood, i.e., $R(\theta,\mathcal{U}) = -\sum_{j\in\mathcal{U}}.\log P(\mathbf{X}^j;\theta)$.
- *SelfT*, (Proposed) Self-training of Section 3.1.
- *MinEnt*, (Proposed) Entropy minimization of Section 3.2.

We will demonstrate the improved prediction performance achieved by the proposed discriminative semi-supervised algorithms (e.g., *SCML-SelfT* or *SCML-MinEnt*) against the baseline generative approaches (e.g., *SCML-MLM*) as well as the supervised methods (e.g., *SCML-Sup*).

In all experiments we make use of the conjugate gradient optimization with the initial iterate set to the *ML-Sup* estimates. The balancing constant λ for the semi-supervised algorithms is chosen by grid search based on the performance on a validation set. The learned LDS models are always assumed to have 1st-order linear dynamics, a possible mismatch with the true dynamics of the data.

4.1. Synthetic robot arm state prediction

We test our approaches on the 2D nonlinear robot arm state estimation task defined in [34]. The original data were constructed for the purpose of *static* regression testing, where the arm-end effector x is determined from two joint angles θ_1 and θ_2 as $x=r_1\cos(\theta_1)+r_2\cos(\theta_1+\theta_2)+\varepsilon$, where $\varepsilon\sim\mathcal{N}(0,\sigma^2)$ for some σ (see Fig. 1(a)). We modify the problem by augmenting the original i.i.d. sampling mechanism for the joint angles $\mathbf{y}=[\theta_1,\theta_2]^\top$ with the 2nd-order temporal dynamics on $\mathbf{y}$'s. More specifically, $\mathbf{y}_t=\mathbf{A}_1\cdot\mathbf{y}_{t-1}+\mathbf{A}_2\cdot(\mathbf{y}_{t-2}-\mathbf{m})+\mathbf{m}+\varepsilon$ for properly chosen $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{m}$, where $\mathbf{y}_t$ is affected by both the rotation of $\mathbf{y}_{t-2}$ and the non-orthonormal transformation of $\mathbf{y}_{t-1}$. Eight sequences of length $\sim$200 were sampled. Fig. 1(b) depicts some example sequence of the 1D measurement $\mathbf{X}$ and the 2D state $\mathbf{Y}$. Note that our LDS model has a suboptimal structure as the true data generating process follows the 2nd-order dynamics with the nonlinear emission function.

Among the eight sequence generated, we randomly choose test, validation, and labeled training sets, each containing one sequence different from one another. For the remaining five sequences, we consider two different unlabeled training sets: $\mathcal{U}_A$ takes three randomly selected sequences, while $\mathcal{U}_B$ has all five sequences. Thus we have a subset relationship $\mathcal{U}_A\subset\mathcal{U}_B$. This partition strategy is repeated five times, and the average results are reported. Table 1 shows the test errors for competing approaches for two unlabeled training sets. The error measures used here is the l_2 error averaged over time slices, $(1/T)\sum_{t=1}^{T}\|\overline{\mathbf{y}}_t-\mathbf{m}_t\|_2$, where $\overline{\mathbf{y}}_t$ is the ground-truth state, and $\mathbf{m}_t$ is the estimated state at time t.

As shown in the table, we again verify that the discriminative learning algorithms for the supervised case (*CML-Sup* and *SCML-Sup*) significantly outperform the generative *ML-Sup*, while the *SCML-Sup* especially performs well demonstrating its robustness

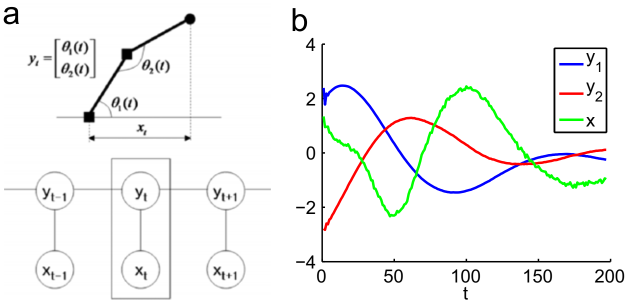

Fig. 1. Robot arm data. (a) Diagram illustrating how robot arm data are generated. (b) Some example samples.

to the suboptimal model structure. Also, when the unlabeled data are provided, the proposed semi-supervised algorithms (*SelfT* and *MinEnt*) yield significant improvement over the supervised learning for all objectives *ML*, *CML*, and *SCML*, than the generative *MLM* can do.

In particular, the *MinEnt* consistently leads to more accurate prediction than the *SelfT*, compensating its higher computational cost. Although both *MinEnt* and *SelfT* shares the underlying assumption that the label prediction on the unlabeled data points, at least some, has to be correct and certain, one possible explanation why *MinEnt* outperforms *SelfT* is that *MinEnt* can deal with small confidences in prediction uncertainty whereas *SelfT* only makes hard decision discarding small possibilities. Such a soft decision by *MinEnt* can be potentially highly effective when the number of unlabeled data points increases. When comparing $\mathcal{U}_A$ and $\mathcal{U}_B$, we see that increasing the number of unlabeled data points can further improve the prediction performance for all approaches.

4.2. Human motion estimation

We next test the proposed methods on the task of 3D body pose estimation from video sequences. The CMU motion capture dataset (http://mocap.cs.cmu.edu/) provides the ground-truth body poses (3D joint angles), which makes it possible to compare competing methods quantitatively. Among the original 59 angles at 31 articulation points, we use $d=39$ by excluding less significant joint angles around fingers and toes as well as those that rarely vary over time. We are particularly interested in two types of motions: walking and golf swing. Both sequences are about 150-frame long with 40 fps rates. The measurement is a 10-dim

Table 1
Test errors in the robot arm dataset.

	ML	*CML*	*SCML*
Sup	1.2506 ± 0.0140	0.9859 ± 0.1153	0.8884 ± 0.0230
MLM			
$\mathcal{U}_A$	1.1814 ± 0.0096	0.9110 ± 0.0421	0.8054 ± 0.0587
$\mathcal{U}_B$	1.0580 ± 0.0074	0.8715 ± 0.0391	0.7915 ± 0.0472
SelfT			
$\mathcal{U}_A$	1.0669 ± 0.0045	0.6512 ± 0.0242	0.5666 ± 0.0271
$\mathcal{U}_B$	0.9963 ± 0.0072	0.6105 ± 0.0521	0.5022 ± 0.0144
MinEnt			
$\mathcal{U}_A$	0.9454 ± 0.0055	0.5311 ± 0.0219	0.4433 ± 0.0274
$\mathcal{U}_B$	0.8952 ± 0.0063	0.5129 ± 0.0112	0.4110 ± 0.0321

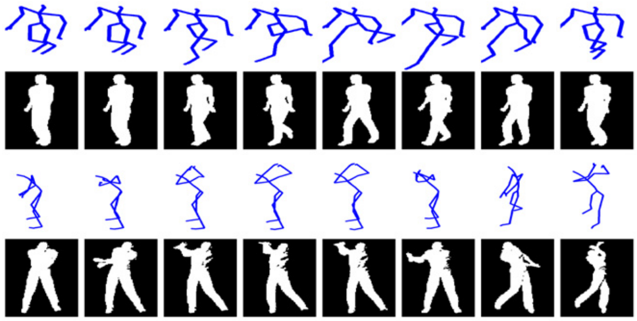

Fig. 2. Selected frames of skeleton and silhouette images for the walking motion (top), and the swing motion (bottom) from the CMU human motion database. The skeleton images are drawn using the ground-truth 3D joint angles, rendered at a particular view point.

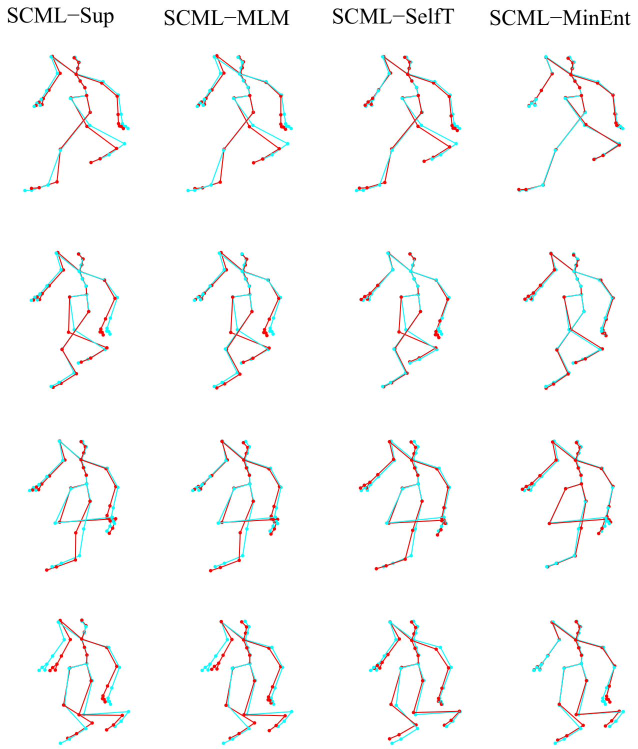

Fig. 3. Prediction results for the walking motion. It highlights four frames at $t=29,58,68,83$ (from top to bottom) predicted by four methods, from left to right, *SCML-Sup*, *SCML-MLM*, *SCML-SelfT*, and *SCML-MinEnt*. In each skeleton figure, the cyan (lighter) is the ground-truth, and the red (darker) is the estimated pose. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Test errors in the CMU human motion dataset.

	Walking			Golf swing		
	ML	*CML*	*SCML*	*ML*	*CML*	*SCML*
Sup	18.60 ± 0.23	16.45 ± 0.12	13.27 ± 0.50	26.28 ± 0.14	19.72 ± 0.15	17.31 ± 0.11
MLM						
$\mathcal{U}_A$	18.28 ± 0.11	16.08 ± 0.11	12.95 ± 0.25	25.54 ± 0.12	19.39 ± 0.17	17.23 ± 0.16
$\mathcal{U}_B$	17.90 ± 0.10	15.46 ± 0.20	11.41 ± 0.15	25.66 ± 0.12	18.69 ± 0.14	16.90 ± 0.17
SelfT						
$\mathcal{U}_A$	17.89 ± 0.20	14.64 ± 0.22	11.37 ± 0.25	22.32 ± 0.09	18.78 ± 0.23	16.89 ± 0.17
$\mathcal{U}_B$	17.25 ± 0.17	14.17 ± 0.21	10.52 ± 0.13	22.47 ± 0.11	17.95 ± 0.21	15.43 ± 0.21
MinEnt						
$\mathcal{U}_A$	17.23 ± 0.17	13.19 ± 0.15	10.30 ± 0.29	22.40 ± 0.13	17.32 ± 0.11	15.69 ± 0.14
$\mathcal{U}_B$	16.05 ± 0.19	12.41 ± 0.27	9.16 ± 0.56	21.94 ± 0.11	17.02 ± 0.12	14.03 ± 0.10

Alt-Moment feature vector extracted from a silhouette image following the method in [35]. The images are taken by a single camera at a fixed view. Fig. 2 shows some selected silhouette images and the corresponding ground-truth body poses visualized using skeletons.

For both motions, we form the semi-supervised setting in the following way. We take 10 sequences (different trials) from one subject where one sequence is held out for testing. The labeled training and the validation sets take one sequence each. For the rest seven sequences, we choose three sequences for the unlabeled training set $\mathcal{U}_A$, and all seven sequences for $\mathcal{U}_B$. Thus $\mathcal{U}_A \subset \mathcal{U}_B$.

We repeat this procedure randomly for five times, and report the l_2 prediction errors for the competing algorithms in Table 2. As before, our discriminative semi-supervised approaches achieve significantly lower estimation errors than the supervised learning and the generative marginal likelihood maximization, where the *MinEnt* yields consistently higher accuracy than the self-training. We also observe that the increased number of unlabeled sequences ($\mathcal{U}_B$) can be useful for further decreasing the motion estimation error. In Figs. 3 and 4, we also visualize the prediction results of the semi-supervised algorithms to illustrate their impact on improving the performance of the most accurate *SCML*-learned LDS.

Our experimental results indicate interesting implication that the unlabeled data, i.e., the silhouette image sequences even without the corresponding human body pose information, can be useful for improving the pose prediction accuracy. One reasonable explanation for this observation is that the proposed semi-supervised learning approaches enforce the LDS models to have higher certainty on the unlabeled data (explicitly done by the *Min-Ent*), which in conjunction with the labeled data, has the effect of enlarging the neighborhood region for reliable model prediction in the measurement space at some extent. So, if the test images are particularly proximal to the silhouette images in the unlabeled data, the prediction performance can be significantly boosted. Another reasonable insight is that our semi-supervised learning algorithms make use of the increased number of image corpus in discriminative manners through conditional models of pose given image, thus yielding more accurate estimators than the standard generative approach based on the image marginal likelihood.

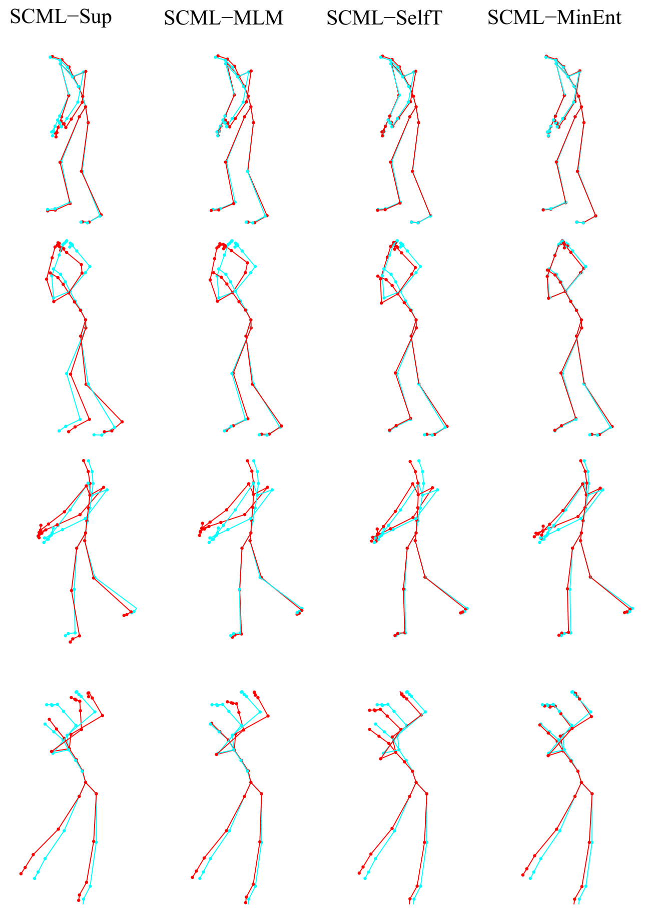

Fig. 4. Prediction results for the golf swing motion. It highlights four frames at t=2,33,64,80 (from top to bottom) predicted by four methods, from left to right, *SCML-Sup*, *SCML-MLM*, *SCML-SelfT*, and *SCML-MinEnt*. In each skeleton figure, the cyan (lighter) is the ground-truth, and the red (darker) is the estimated pose. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Concluding remarks

We introduced novel discriminative semi-supervised learning algorithms for dynamical systems. Framed in the unifying semi-supervised framework, the proposed approaches can exploit a large amount of unlabeled video data in a discriminative manner to improve the prediction performance significantly. Despite the outstanding performance, the nested recursion in the entropy minimization method can be computationally demanding especially when the state dimension is high. This issue may be potentially addressed by two different ways: either incorporating efficient, possibly approximated, inference algorithms such as the importance sampling, or reducing the state dimension by discriminative dimensionality reduction methods such as the canonical correlation analysis. We will leave them as future work.

Appendix A. Derivation for Eq. (11)

We derive (11) as follows:

$$\frac{\partial}{\partial\theta}\int_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X})\log P(\mathbf{Y}|\mathbf{X})\,d\mathbf{Y} = \frac{\partial}{\partial\theta}\int_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X})\log P(\mathbf{X},\mathbf{Y})\,d\mathbf{Y} - \frac{\partial}{\partial\theta}\log P(\mathbf{X}) \tag{26}$$

$$= \int_{\mathbf{Y}}\left(\frac{\partial}{\partial\theta}P(\mathbf{Y}|\mathbf{X})\right)\log P(\mathbf{X},\mathbf{Y})\,d\mathbf{Y} \tag{27}$$

$$+\int_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X})\left(\frac{\partial}{\partial\theta}\log P(\mathbf{X},\mathbf{Y})\right)d\mathbf{Y} - \frac{\partial}{\partial\theta}\log P(\mathbf{X})$$

$$= \int_{\mathbf{Y}}\left(\frac{\partial}{\partial\theta}P(\mathbf{Y}|\mathbf{X})\right)\log P(\mathbf{X},\mathbf{Y})\,d\mathbf{Y} \tag{28}$$

$$= \int_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X})\left(\frac{\partial}{\partial\theta}\log P(\mathbf{Y}|\mathbf{X})\right)\log P(\mathbf{X},\mathbf{Y})\,d\mathbf{Y} \tag{29}$$

$$= \mathbb{E}\left[\log P(\mathbf{X},\mathbf{Y})\cdot\frac{\partial}{\partial\theta}\log P(\mathbf{X},\mathbf{Y})\right] - \mathbb{E}[\log P(\mathbf{X},\mathbf{Y})]\cdot\mathbb{E}\left[\frac{\partial}{\partial\theta}\log P(\mathbf{X},\mathbf{Y})\right]. \tag{30}$$

References

[1] M. Kim, V. Pavlovic, Discriminative learning of dynamical systems for motion tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[2] V. Pavlovic, J.M. Rehg, J. MacCormick, Learning switching linear models of human motion, in: Proceedings of the Advances in Neural Information Processing Systems, 2000.

[3] P. Smyth, Belief networks, hidden Markov models, and Markov random fields: a unifying view, Pattern Recognition Letters 18 (11–13) (1997) 1261–1268.

[4] J.C. Nascimento, J.S. Marques, Improving the robustness of parametric shape tracking with switched multiple models, Pattern Recognition 35 (12) (2002) 2711–2718.

[5] J. Frankel, S. King, Factoring Gaussian precision matrices for linear dynamic models, Pattern Recognition Letters 28 (16) (2007) 2264–2272.

[6] A. Garzelli, F. Nencini, Panchromatic sharpening of remote sensing images using a multiscale Kalman filter, Pattern Recognition 40 (12) (2007) 3568–3577.

[7] Z. Ghahramani, S. Roweis, Learning nonlinear dynamical systems using an EM algorithm, in: Proceedings of the Advances in Neural Information Processing Systems, 1999.

[8] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, Robust online appearance models for visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2001) 1296–1311.

[9] D. Ross, S. Osindero, R. Zemel, Combining discriminative features to infer complex trajectories, in: International Conference on Machine Learning, 2006.

[10] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3D human motion estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[11] L. Taycher, D. Demirdjian, T. Darrell, G. Shakhnarovich, Conditional random people: tracking humans with CRFs and grid filters, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[12] M. Kim, V. Pavlovic, Discriminative learning for dynamic state prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1847–1861.

[13] P. Woodland, D. Povey, Large scale discriminative training of hidden Markov models for speech recognition, Computer Speech and Language 16 (1) (2002) 25–47.

[14] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, 2001.

[15] X. Zhu, A.B. Goldberg, Introduction to Semi-Supervised Learning, Morgan & Claypool, 2009.

[16] V. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.

[17] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, in: Proceedings of the Advances in Neural Information Processing Systems, 2004.

[18] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, D. Schuurmans, Semi-supervised conditional random fields for improved sequence segmentation and labeling, in: The 44th Annual Meeting of the Association for Computational Linguistics (ACL), 2006.

[19] C.-H. Lee, S. Wang, F. Jiao, D. Schuurmans, R. Greiner, Learning to model spatial dependency: semi-supervised discriminative random fields, in: Proceedings of the Advances in Neural Information Processing Systems, 2006.

[20] B. North, M.I.A. Blake, J. Rittscher, Learning and classification of complex dynamics, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2000) 1016–1034.

[21] P. Abbeel, A. Coates, M. Montemerlo, A.Y. Ng, S. Thrun, Discriminative training of Kalman filters, in: Proceedings of Robotics, 2005.

[22] N.D. Lawrence, Gaussian process models for visualisation of high dimensional data, in: Neural Information Processing Systems (NIPS), 2003.

[23] J.M. Wang, D.J. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 283–298.

[24] J. Pang, L. Qing, Q. Huang, S. Jiang, W. Gao, Monocular tracking 3D people by Gaussian process spatio-temporal variable model, in: IEEE International Conference on Image Processing (ICIP), 2007.

[25] A. Kanaujia, C. Sminchisescu, D. Metaxas, Semi-supervised hierarchical models for 3D human pose reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[26] E. Riloff, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the 7th Conference on Natural Language Learning (CoNLL), 2003.

[27] B. Maeireizo, D. Litman, R. Hwa, Co-training for predicting emotions with spoken dialogue data, in: The Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), 2004.

[28] C. Rosenberg, M. Hebert, H. Schneiderman, Semi-supervised self-training of object detection models, in: 7th IEEE Workshop on Applications of Computer Vision, 2005.

[29] Y. Bar-Shalom, X.-R. Li, Estimation and Tracking: Principles, Techniques, and Software, Artech House, Boston, 1993.

[30] S. Kakade, Y. Teh, S. Roweis, An alternate objective function for Markovian fields, in: International Conference on Machine Learning, 2002.

[31] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B 39 (1) (1977) 1–38.

[32] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons Inc, 1973.

[33] M.I. Jordan, Graphical models, Statistical Science 19 (2004) 140–155.

[34] D.J.C. MacKay, A practical Bayesian framework for backpropagation networks, Neural Computation 4 (1992) 448–472.

[35] T.-P. Tian, R. Li, S. Sclaroff, Articulated pose estimation in a learned smooth space of feasible solutions, in: Proceedings of IEEE Workshop in Computer Vision and Pattern Recognition, 2005.

Minyoung Kim received the BS and MS degrees both in Computer Science and Engineering in Seoul National University, South Korea. He earned the PhD degree in Computer Science from Rutgers University in 2008. From 2009 to 2010 he was a postdoctoral researcher at the Robotics Institute of Carnegie Mellon University. He is currently an Assistant Professor in the Department of Electronic and Information Engineering at Seoul National University of Science and Technology in Korea. His primary research interest is machine learning and computer vision. His research focus includes graphical models, motion estimation/tracking, discriminative models/learning, kernel methods, and dimensionality reduction.