A Linear Dynamical System Model for Text

David Belanger
College of Information and Computer Sciences, University of Massachusetts Amherst

Sham Kakade
Microsoft Research

Abstract

Low dimensional representations of words allow accurate NLP models to be trained on limited annotated data. While most representations ignore words’ local context, a natural way to induce context-dependent representations is to perform inference in a probabilistic latent-variable sequence model. Given the recent success of continuous vector space word representations, we provide such an inference procedure for continuous states, where words’ representations are given by the posterior mean of a linear dynamical system. Here, efficient inference can be performed using Kalman filtering. Our learning algorithm is extremely scalable, operating on simple cooccurrence counts for both parameter initialization using the method of moments and subsequent iterations of EM. In our experiments, we employ our inferred word embeddings as features in standard tagging tasks, obtaining significant accuracy improvements. Finally, the Kalman filter updates can be seen as a linear recurrent neural network. We demonstrate that using the parameters of our model to initialize a non-linear recurrent neural network language model reduces its training time by a day and yields lower perplexity.

1. Introduction

In many NLP applications, there is limited available labeled training data, but tremendous quantities of unlabeled, in-domain text. An effective semi-supervised learning technique is to learn word embeddings on the unlabeled data, which map every word to a low dimensional dense vector (Bengio et al., 2006; Mikolov et al., 2013; Pennington et al., 2014), and then use these as features for supervised training on the labeled data (Turian et al., 2010; Paslsos et al., 2014; Bansal et al., 2014). Furthermore in many deep architectures for NLP, the first layer maps words to low-dimensional vectors, and these parameters are initialized with unsupervised embeddings (Collobert et al., 2011; Socher et al., 2013; Vinyals et al., 2014).

Most of these methods embed word types, i.e., words independent of local context, as opposed to work tokens, i.e., instances of words within their context. Ideally we would have a different representation per token. For example, depending on the context, “bank” is the side of a river or a financial institution. Furthermore, we would like such embeddings to come from a probabilistic sequence model that allows us to study the transition dynamics of text generation in low dimensional space.

We present a method for obtaining such context-dependent token embeddings, using a generative model with a vector-valued latent variable per token and performing posterior inference for each sentence. Specifically, we employ a Gaussian linear dynamical system (LDS), with efficient inference from a Kalman filter. To learn the LDS parameters, we use a two-stage procedure, initializing with the method of moments, and then performing EM with the approximate second order statistics (ASOS) technique of Martens (2010). Overall, after taking a single pass over the training corpus, the runtime of our approximate maximum-likelihood estimation (MLE) procedure is independent of the amount of training data since it operates on aggregate co-occurrence counts. Furthermore, performing inference to obtain token embeddings has the same time complexity as widely-used discrete first-order sequence models.

We fit the LDS to a one-hot encoding of each token in the input text sequence. Therefore, the LDS is a mis-specified generative model, since draws from it are not proper indicator vectors. However, we embrace this multivariate Gaussian model instead of a continuous-state dynamical system with a multinomial link function because the Gaussian LDS offers several desirable scalability properties: (1) Kalman
filter of our learning iterations does not scale with the corpus size, (3) we can initialize EM using a method-of-moments estimator that requires a single SVD of a co-occurrence matrix, (4) our M-step updates are simple least-squares problems, solvable in closed form, (4) if we had used a multinomial link function, we would have performed inference using extended Kalman filtering, which makes a second-order approximation of the log-likelihood, and thus leads to a Gaussian LDS anyway (Ghahramani & Roweis, 1999), and (5) by using EM, we avoid stochastic-gradient-based optimization, which requires careful tuning for nonconvex problems. A naive application of our method scales to large amounts of training data, but not high-dimensional observations. In response, the paper contributes a variety of novel methods for scaling up our learning techniques to handle large input vocabularies.

We employ our inferred token embeddings as features for part of speech (POS) and named entity recognition (NER) taggers. For POS, we obtain a 30% relative error reduction when applying a local classifier to our context-dependent embeddings rather than Word2Vec context-independent embeddings (Mikolov et al., 2013). When using our token embeddings as additional features in lexicalized POS and NER taggers, which already have explicit features and test-time inference for context-dependence, we obtain significant gains over the baseline, performing as well as using Word2Vec embeddings. We also present experiments demonstrating that the transition dynamics of the LDS capture salient patterns, such as transforming first names into last names.

Finally, the functional form of the Kalman filter update equations for our LDS are identical to the updates of a recurrent neural network (RNN) language model without non-linearities (Mikolov, 2012). A key difference between the LDS and an RNN, however, is that the LDS provides a natural backwards pass, using Kalman smoothing, where a token’s embedding depends on text to both the right and left. Drawing on the parallelism between filtering and the RNN, we use the LDS parameters, which can be estimated very quickly using our techniques, to initialize gradient-based optimization of a nonlinear RNN. This yields a significant decrease in perplexity vs. the baseline RNN, and only requires 70% as many training epochs, saving 1 day on a single CPU core.

2. Related Work

We provide a continuous analog of popular discrete-state generative models used in NLP for inducing class membership for tokens, including class-based language models (Brown et al., 1992; Chelba & Jelinek, 2000) and induction of POS tags (Christodoulopoulos et al., 2010). In particular, Brown clusters (Brown et al., 1992) are commonly used by practitioners with lots of unlabeled in-domain data.

Our learning algorithm is very scalable because it operates on aggregate count matrices, rather than individual tokens. Similar algorithms have been proposed for obtaining type-level embeddings via matrix factorization (Pennington et al., 2014; Levy & Goldberg, 2014). However, these are context-independent and ignore the transition dynamics that link tokens’ embeddings. Furthermore, they require careful tuning of stochastic gradient methods. Previous methods for token-level embeddings either use a rigid set of prototypes (Huang et al., 2012; Neelakantan et al., 2014) or embed the token’s context, ignoring the token itself (Dhillon et al., 2011).

For learning discrete-state latent variable models, spectral learning methods also use count matrices, and thus are similarly scalable (Anandkumar et al., 2014). However, an LDS offers key advantages: we do not use third-order moments, which are difficult to estimate, and we perform approximate MLE, rather than the method of moments, which exhibits poor statistical efficiency.

Recently, RNNs have been used to provide impressive results in NLP applications including translation (Sutskever et al., 2014), language modeling (Mikolov et al., 2015), and parsing (Vinyals et al., 2014). We do not attempt to replace these with a Kalman filter, as we expect non-linearities are crucial for capturing long-term interactions and rigid, combinatorial constraints in the outputs. However, RNN training can take days, even on GPUs, and requires careful tuning of stochastic gradient step sizes. Given the scalability of our parameter-free training algorithm, and our favorable preliminary results using the LDS to initialize a nonlinear RNN, we encourage further work on using linear latent variable models and the Gaussian approximations of multinomial data to develop sophisticated initialization methods. Already, practitioners have started using such techniques for initializing simple nonlinear deep neural networks using the recommendations of Saxe et al. (2014). Finally, our work differs from Pasa & Sperduti (2014), who initialize an RNN using spectral techniques, in that we perform maximum-likelihood learning. We found this crucial for good performance in our NLP experiments.

3. Background: Gaussian Linear Dynamical Systems

We consider sequences of observations \( w_1, \ldots, w_n \), where each \( w_t \) is a \( V \)-dimensional vector. A Gaussian LDS follows the following generative model (Kalman, 1960; Roweis & Ghahramani, 1999):

\[
x_t = Ax_{t-1} + \eta \tag{1}
\]

\[
w_t = Cx_t + \epsilon, \tag{2}
\]
where \( h < V \) is the dimensionality of the hidden states \( x_t \) and \( \epsilon \sim N(0, D) \) and \( \eta \sim N(0, Q) \). For simplicity, we assume \( x_0 \) is constant.

The latent space for \( x \) is completely unobserved and we could choose any coordinate system for it while maintaining the same likelihood value. Therefore, without loss of generality, we can either fix \( A = I \) or \( Q = I \), and we fix \( Q \). Furthermore, note that the magnitude of the maximum eigenvalue of \( A \) must be no larger than 1 if the system is stable. We assume that the data we fit to has been centered, in which case the maximum eigenvalue is strictly less than 1, since this implies \( x_t \) is asymptotically mean zero (independent of \( x_0 \)), so that \( x_t \) is also asymptotically mean zero.

Finally, define the covariance at lag \( k \) to be

\[
\Psi_k = \mathbb{E}[w_{t+k}w_t^T],
\] (3)

which is valid because we assume the data to be mean zero. Our learning algorithms require only a few \( \Psi_k \) (up to about \( k = 10 \) in practice) as input. These matrices can be gathered using a single pass over the data, and their size does not depend on the amount of data. Furthermore, constructing these matrices can be accelerated by splitting the data into chunks, and aggregating separate matrices afterwards.

3.1. Inference

The \( x_t \) are distributed as a multivariate Gaussian under both the LDS prior and posterior (conditional on observations \( w \)), so they can be fully characterized by a mean and covariance. We use \( \hat{x}_t \) and \( S_t \) for the mean and covariance under the posterior for \( x_t \) given \( w_{1:t-1} \), computed using Kalman filtering, and \( \hat{x}_t \) and \( S_T \) when considering the posterior for \( x_t \) given all the data \( w_{1:T} \), computed using Kalman smoothing. In Appendix B.1 we provide the full filtering and smoothing updates, which compute different means and variances for every timestep. Note that the updates require inverting a \( V \times V \) matrix in every step.

We employ the widely-used ‘steady-state’ approximation, which yields substantially more efficient filtering and smoothing updates (Rugh, 1996). A key property of filtering and smoothing is that the updates to \( S_t \) and \( S_T \) do not depend on the actual observations, but only on the model’s parameters. Furthermore, they will converge quickly to time-independent ‘steady-state’ values. Define \( S_1 = \mathbb{E}\left[\hat{x}_t\hat{x}_t^T\mid w_{1:t-1}\right] \) to be the asymptotic limit of the covariance \( S_t \) under the posterior for each \( x_t \) given its history (at steady state, this is shared for all \( t \)). Here, expectation is taken with respect to both time and the posterior for the latent variables. This satisfies

\[
S_1 = AS_1A^T + Q,
\]

which can be solved for quickly using fixed point iteration. Similarly, we can solve for \( \Sigma_0 = \mathbb{E}\left[\hat{x}_t\hat{x}_t^T\mid w_{1:1}\right] \). Note that steady state, a property of the posterior, is unrelated to the stationary distribution of the LDS, which is unconditional on observations.

Under the steady state assumption, we can perform filtering and smoothing using substantially more efficient updates. We have:

\[
\hat{x}_t = (A - KA)\hat{x}_{t-1} + Kw_t \quad (4)
\]
\[
\bar{x}_t = J\hat{x}_{t+1} + (I - JA)\hat{x}_t \quad (5)
\]

Here, the steady-state Kalman gain matrix is:

\[
K = \Sigma_1C^T S_{ss}^{-1} \in \mathbb{R}^{h \times V},
\] (6)

where we define

\[
S_{ss} = CS_1C^T + D,
\] (7)

the unconditional prior covariance for \( w \) under the model. Note that \((A - KA)\) is data-independent and can be pre-computed, as can the smoothing matrix \( J = \Sigma_0A^T (\Sigma_1)^{-1} \). For long sequences, steady-state filtering provides asymptotically exact inference. However, for short sequences it is an approximation.

3.2. Learning: Expectation-Maximization

See Ghahramani & Hinton (1996) for a full exposition on learning the parameters of an LDS using EM. Under the steady-state assumption, the M step requires:

\[
\mathbb{E}[\hat{x}_t\hat{x}_t^T], \quad \mathbb{E}[\hat{x}_t\hat{x}_{t+1}^T], \quad \mathbb{E}[\hat{x}_t w_t^T], \quad (8)
\]

where the expectation is taken with respect to time and the posterior for the latent variables. This can be computed using Kalman smoothing and then averaging over time. The M step can then be done in closed form, since it is solving least-squares regressions for \( x_{t+1} \) against \( x_t \) and \( w_t \) against \( x_t \) to obtain \( A \) and \( C \). Lastly, \( D \) can be recovered using:

\[
D = \Psi_0 - CE\left[\hat{x}_t w_t^T\right] - \mathbb{E}\left[w_t\hat{x}_t^T\right] C^T + CE\left[\hat{x}\hat{x}_t^T\right] C^T \quad (9)
\]

3.3. Learning: EM with ASOS (Martens, 2010)

EM requires recomputing the second order statistics (8) in every iteration. While these can be computed using Kalman smoothing on the entire training set, we are interested in datasets with billions of timesteps. Fortunately, we can avoid smoothing by employing the ASOS (approximate second order statistics) method of Martens (2010), which directly performs inference about the time-averaged second order statistics.

Under the steady-state assumption, this is doable because we can recursively define relationships between second order statistics at lag \( k \) and at lag \( k + 1 \) using the recursive
relationships of the underlying dynamical system. Namely, rather than performing posterior inference by recursively applying the linear operations (4) and (5), and then averaging over time, we switch the order of these operations and apply the linear operators to time-averaged second order statistics. For example, the following equality is an immediate consequence of the filtering equation (4) (where expectation is with respect to t and the posterior for x):

\[
E[\hat{x}_t^2 w_t^T] = (A - KCA)E[\hat{x}_{t-1}^2 w_t^T] + KE[w_tw_t^T] \quad (10)
\]

ASOS uses a number of such recursions, along with methods for estimating covariances at a time horizon \(r\). These covariances can be approximated by assuming that they are exactly described by the current estimate of the model parameters. Therefore, unlike standard EM, performing EM with ASOS allows us to precompute an empirical estimate of the \(\hat{\Psi}_k\) at various lags (up to about \(r = 10\)) and then never touch the data again. Furthermore, (Martens, 2010) demonstrates that the ASOS approximation is consistent. Namely, the error in approximating the time-averaged second order statistics vanishes with infinite data when evaluated at the MLE parameters. Overall, ASOS scales linearly with \(r\) and the cost of multiplying by the \(\hat{\Psi}_k\).

### 3.4. Learning: Subspace Identification

We initialize EM using Subspace Identification (SSID), a family of method-of-moments estimators that use spectral decomposition to recover LDS parameters (Van Overschee & De Moor, 1996). The rationale for such a combination is that the method of moments is statistically consistent, so performing it on reasonably-sized datasets will yield parameters in the neighborhood of the global optimum, and then EM will perform local hill climbing to find a local optimum of the marginal likelihood. For LDS, this combination yields empirical accuracy gains in Smith et al. (1999). A related two-stage estimator, where the local search of EM is replaced with a single Newton step on the local likelihood surface, is known to be minimax optimal, under certain local asymptotic normality conditions (Le Cam, 1974).

For our particular application, we use SSID as an approximate method, where it is not statistically consistent, due to the mis-specification of fitting indicator-vector data as a multivariate Gaussian. In our experiments, we discuss the superiority of SSID+EM rather than just SSID. We do not present results using EM initialized randomly rather than with SSID, since we found it very difficult for our high dimensional problems to generate initial parameters that allowed EM to reach high likelihoods.

We employ the ‘n4sid’ algorithm of Van Overschee & De Moor (1994). Define \(r\) to be a small integer. Define the \((rV) \times h\) matrix \(\Gamma_r = [C; CA; CA^2; \ldots; CA^{r-1}]\), where ‘;’ denotes vertical concatenation. Also define the x-w covariance at lag 0: \(G = E[\hat{x}_{t+1}^T w_t]\) and the \(h \times (rV)\) matrix \(\Delta_r = [A^{r-1}G \ A^{r-2}G \ldots AG G]\).

Next, define the Hankel matrix

\[
H_r = \begin{pmatrix}
\Psi_r & \Psi_{r-1} & \Psi_{r-2} & \ldots & \Psi_1 \\
\Psi_{r+1} & \Psi_r & \Psi_{r-1} & \ldots & \Psi_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\Psi_{2r-1} & \Psi_{2r-2} & \Psi_{r-3} & \ldots & \Psi_r
\end{pmatrix} \quad (11)
\]

Then, we have \(H_r = \Gamma_r \Delta_r\).

Let \((U, S, V)\) result from a rank-\(h\) SVD of an empirical estimate of \(H_r\), from which we set \(\Gamma_r = US^{2r/2}\) and \(\Delta_r = S^{2r/2}V^T\). To recover the LDS parameters, we first define \(\Delta_r^{1/2}(r-1)\) to be the submatrix of \(\Delta_r\) corresponding to the first \((r-1)\) blocks. Similarly, define \(\Delta_r^{2r-r}\). From the definition of \(\Delta_r\), we have that \(A\Delta_r^{2r-r} = \Delta_r^{1/2}(r-1)\), so we can estimate \(A\) as \(A = \Delta_r^{1/2}(r-1)(\Delta_r^{2r-r})^{+}\). Next, one can read off an estimate for \(C\) as the first block of \(\Gamma_r\). Alternatively, since the previous step gives us a value for \(A\) one can set up a regression problem similar to the previous step to solve for \(C\) by invoking the block structure in \(\Gamma_r\).

Finally, we need to recover the covariance matrix \(D\). We first find the asymptotic latent covariance \(\Sigma_1\) using fixed-point iteration \(\Sigma_1 = A\Sigma_1A^T + Q\). From this, we set \(D\) using a similar update as (12), which uses statistics of the LDS posterior, except here \(\Sigma_1\) is unconditional on data and is purely a function of the LDS parameters.

\[
D = \Psi_0 - C\Sigma_1C^T. \quad (12)
\]

### 4. Linear Dynamical Systems for Text

We fit an LDS to text using SSID to initialize EM, where the E step is performed using ASOS. A summary of the procedure is provided in Algorithm 1. SSID and ASOS scale to extremely large training sets, since they only require the \(\hat{\Psi}_k\) matrices, for small \(k\). However, they cannot directly handle the very high dimensionality of text observations (vocabulary size \(V \approx 10^8\)). In this section, we first
describe particular properties of the data distribution. Then, we describe novel techniques for leveraging these properties to yield scalable learning algorithms.

Define \( \tilde{\mathbf{w}}_t \) as an indicator vector that is 1 in the index of the word at time \( t \) and define \( \mu_t \) to be the corpus frequency of word type \( i \). We fit to the mean-zero observations \( \mathbf{w}_t = \tilde{\mathbf{w}}_t - \mu \). Note that the LDS will not generate observations with the structure of a one-hot vector shifted by a constant mean, so we cannot use it directly as a generative language model. On the other hand, we can still fit models to training data with this structure, perform posterior inference given observations, assess the likelihood of a corpus, etc. In our experiments, we demonstrate the usefulness of these in a variety of applications. We have:

\[
\Psi_0 = \mathbb{E}[\mathbf{w}_t \mathbf{w}_t^\top] = \mathbb{E}[\tilde{\mathbf{w}}_t \tilde{\mathbf{w}}_t^\top] - \mu \mu^\top = \text{diag}(\mu) - \mu \mu^\top, \tag{13}
\]

while at higher lags,

\[
\Psi_k = \mathbb{E}[\mathbf{w}_t \mathbf{w}_t^\top] = \mathbb{E}[\tilde{\mathbf{w}}_t \tilde{\mathbf{w}}_t^\top] - \mu \mu^\top. \tag{14}
\]

Approximating these covariances from a length-\( T \) corpus:

\[
\mu_t = \mathbb{E}[\tilde{\mathbf{w}}_t] = \frac{1}{T} \mathbb{E}[\text{#(word } i \text{ appears)}], \tag{15}
\]

where \( \#() \) denotes the count of an event. We also have

\[
\mathbb{E}[\tilde{\mathbf{w}}_t \tilde{\mathbf{w}}_t^\top]_{i,j} = \frac{1}{T} \mathbb{E}[\text{#(word } i \text{ appears with word } j \text{ } k \text{ positions to the right)}]. \tag{16}
\]

For real-world data, (16) will be extremely sparse, with the number of nonzeros substantially less than both \( V^2 \) and the length of the corpus. The fact that (14) is sparse-minus-low-rank and (13) is diagonal-minus-low-rank is critical for scaling up the learning algorithms. First of all, we do not instantiate these as \( V \times V \) dense matrices, but operate directly on their factorized structure. Second, in Sec. 4.2 we show how the structure of (13) allows us to model full-rank \( V \times V \) noise covariance matrices implicitly. Strictly speaking, the number of nonzeros in (16) will increase as the corpus size increases, due to heavy-tailed word co-occurrence statistics. However, this growth is sublinear in \( T \) and can be mitigated by ignoring rare words.

Unfortunately, each \( \Psi_k \) is rank-deficient. Not only is \( \mathbb{E}[\mathbf{w}_t] = 0 \), but also the sum of every \( \mathbf{w}_t \) is zero (because \( \tilde{\mathbf{w}}_t \)) is a one-hot vector and \( \mu_t \) is a vector of word frequencies). Define \( \mathbf{1} \) to be the length-\( V \) vector of ones. Then, our data lives on \( \mathbf{1}^\perp \), the \( d - 1 \) dimensional subspace, orthogonal to \( \mathbf{1} \). Doing maximum likelihood in \( \mathbb{R}^V \) instead of \( \mathbf{1}^\perp \) will lead to a degenerate likelihood function, since the empirical variance in the \( \mathbf{1} \) direction is 0. However, projecting the data to this subspace breaks the special structure described above, so we instead work in \( \mathbb{R}^V \) and perform projections onto \( \mathbf{1}^\perp \) implicitly as-needed. Fortunately, both SSID and EM find \( \mathbf{C} \) that lies in the column space of the data, so iterations of our learning algorithm will maintain that \( \mathbf{1} \notin \text{col}(\mathbf{C}) \). In Appendix A.2, we describe how to handle this rank deficiency when computing the Kalman gain.

Note that we could have used pretrained type-level embeddings to project our corpus and then train an LDS on low-dimensional dense observations. However, this is vulnerable to the subspace of the type-level embeddings, which are not trained to maximize the likelihood of a sequence model, and thus might not capture proper syntactic and semantic information. We will release the code of our implementation. SSID requires simple scripting on top of a sparse linear algebra library. Our EM implementation consists of small modifications to Martens’ public ASOS code.

### 4.1. Scalable Spectral Decomposition

SSID requires a rank-\( h \) SVD of the very large block Hankel matrix \( H_r \) (11). We employ the randomized approximate SVD algorithm of Halko et al. (2011). To factorize a matrix \( X \), this requires repeated multiplication by \( X \) and by \( X^\top \). All the submatrices in \( H_r \) are sparse-minus-low-rank, so we handle the sparse and low-rank terms individually within the multiplication subroutines.

### 4.2. Modeling Full-Rank Noise Covariance

The noise covariance matrix \( D \) is \( V \times V \), which is unmanageably large for our application, and thus it is reasonable to employ a spherical \( D = dI \) or diagonal \( D = \text{diag}(d_1, \ldots, d_V) \) approximation. For our problem, however, we found that these approximations performed poorly. Because of the property \( \mathbf{1}^\top \mathbf{u}_t = 0 \), off-diagonal elements of \( D \) are critical for modeling the anti-correlations between coordinates. This would have been captured if we passed \( \mathbf{u}_t \) through a logistic multinomial link function. However, this prevents simple inference using Kalman filter. To maintain conjugacy, practitioners sometimes employ the quadratic upper bound to a logistic multinomial likelihood introduced in Böhning (1992), which hard-codes the coordinate-wise anticorrelations via

\[
D = \frac{1}{2} \left[ I - \frac{1}{V+1} \mathbf{1} \mathbf{1}^\top \right].
\]

However, we found this data-independent estimator performed poorly.

Instead, we exploit a particular property of the SSID and EM estimators for \( D \) in (12) and (9). Namely, both set \( D \) to \( \Psi_0 \) minus a low-rank matrix, and thus \( D \) is diagonal-minus-low-rank, due to the structure in (13). For the LDS, we mostly seek to manipulate the precision matrix \( D^{-1} \). While instantiating this dense \( V \times V \) matrix is infeasible, multiplication by \( D^{-1} \) and evaluation of \( \det(D^{-1}) \) can both
be done efficiently using the ShermanWoodbury-Morrison formula (Appendix B.2). In Appendix A.3, we also leverage the formula to efficiently evaluate the training likelihood. These uses of the formula differ from its common usage for LDS, when not using the steady-state assumption and the posterior precision matrix needs to be updated using rank one updates to the covariance. Our technique is particular to fitting indicator-vector data as a multivariate Gaussian.

4.3. Whitening

Before applying our learning algorithms, we first whiten the $\Psi$ matrices with the diagonal transformation.

$$W = \Psi_0^{-\frac{1}{2}} = \text{diag}(\mu_1^{-\frac{1}{2}}, \ldots, \mu_v^{-\frac{1}{2}}).$$

Fitting to $W\Psi_k W^T$, rather than $\Psi_k$, maintains the data’s sparse-minus-low-rank and diagonal-minus-low-rank structures. Furthermore, EM is unaffected, i.e., applying EM to linearly-transformed data is equivalent to learning on the original data and then transforming post-hoc.

On the other hand, the SSID output is affected by whitening, since the squared reconstruction loss that SVD implicitly minimizes depends on the coordinate system of the data. We found such whitening crucial for obtaining high-quality initial parameters. Whitening for SSID, which is recommended by Van Overschee & De Moor (1996), solves a very similar factorization problem as canonical correlation analysis between words and their contexts, which has been used successfully to learn word embeddings (Dhillon et al., 2011; 2012) and identify the parameters of class-based language models (Stratos et al., 2014).

In Appendix A.1 we also provide an algorithm, which relies on whitening, for manually ensuring the $D$ returned by SSID is PSD, without needing to factorize a $V \times V$ matrix. Such manual correction is unnecessary during EM, since the estimator (9) is guaranteed to be PSD.

5. Embedding Tokens using the LDS

The only data-dependent term in the steady-state filtering and smoothing equations (4) and (5) is $Kw_t$. Since $w_t$ can take on only $V$ possible values, we precompute these word-type-level vectors. The computational cost of filtering/smoothing a length $T$ sequence is $O(Th^2)$, which is identical to the cost of inference on a discrete first-order sequence model. (6) is not directly usable to obtain $K$, due to the data’s rank-deficiency, and we provide an efficient alternative in Appendix A.2. This also requires the matrix inversion lemma to avoid instantiating $S^{-1}$ in (6).

In our experiments we use the latent space to define features for tokens. However, distances in this space are not well-defined, since the likelihood is invariant to any linear transformation of the latent variables. To place $x_t$ in reasonable coordinates, we compute the empirical posterior covariance $M = E[xT x^T]$ on the training data (using ASOS). Then, we whiten $x_t$ using $M^{-\frac{1}{2}}$ and project the result onto the unit sphere.

6. Relation to Recurrent Neural Networks

We now highlight the similarity between the parametrization of an RNN architecture commonly used for language modeling and our Kalman filter. This allows us to use our LDS as a novel method for initializing the parameters of a non-linear RNN, which we explore in Sec. 7.4. Following (Mikolov, 2012) we consider the network structure:

$$h_t = \sigma(Ah_{t-1} + Bw_{t-1})$$
$$w_t \sim \text{SoftMax}(Ch_t),$$

Here, we employ the SoftMax transformation of a vector $v$ as $v_i \rightarrow \exp(v_i)/\sum_k \exp(v_k)$. The coordinate-wise non-linearity $\sigma(\cdot)$ is, for example, a sigmoid, and the network is initialized with some fixed vector $h_0$.

Consider the use of the steady-state Kalman filter (4) as an online predictor, where the mean prediction for $w_t$ is given by $C\hat{x}_t$. Then, if we replace $\sigma$ and SoftMax with the identity, the Kalman filter and the RNN have the same set of parameters, where we $B$ corresponds to $K$ and $A$ corresponds to $(A - KCA)$. In terms of the state dynamics, the LDS may provide parameters that are reasonable for a nonlinear RNN, since the sigmoid $\sigma$ has a regime for inputs close to zero where it behaves like the identity. A linear approximation of SoftMax$(\cdot)$ ignores mutual exclusivity. However, we discuss in Section 4.2 that using a full-rank $D$ captures some coordinate-wise anti-correlations. Also, (19) does not affect the state evolution in (18).

A key difference between the LDS and the RNN is that the LDS provides a backwards pass, using Kalman smoothing, where $\hat{x}_t$ depends on words to the right. For RNNs, this would requires separate model (Schuster & Paliwal, 1997).

7. Experiments

7.1. LDS Transition Dynamics

Many popular word embedding methods learn word-to-vector mappings, but do not learn the dynamics of text’s evolution in the latent space. Using the specific LDS model we describe in the next section, we employ the transition matrix $A$ to explore properties of these dynamics. Because the state evolution is linear, it can be studied easily using a spectral decomposition. Namely, $A$ converts its left singular vectors into (scaled) right singular vectors. For each vector, we find the words most likely to be generated from
The accuracy of a local classifier on the PTB dev set. The LDS hyperparameters were selected by maximizing for each type, by adding frequent types with “OOV.” We employ \( r = 4 \) for SSID, \( r = 7 \) for EM, and \( h = 200 \). We add 1000 pseudocounts for each type, by adding \( \frac{1000}{r} \) to each coordinate of \( \mu \).

The LDS hyperparameters were selected by maximizing the accuracy of a local classifier on the PTB dev set.

<table>
<thead>
<tr>
<th>Right Singular Vector</th>
<th>Left Singular Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>islamist lebanese israeli</td>
<td>territories immigrants sea</td>
</tr>
<tr>
<td>palestinian british latin</td>
<td>films communities nationals</td>
</tr>
<tr>
<td>japanese shiite greek</td>
<td>rivals africa clients</td>
</tr>
<tr>
<td>chris mike steve</td>
<td>evans anderson harris</td>
</tr>
<tr>
<td>jason tim jeff</td>
<td>robinson smith phillips</td>
</tr>
<tr>
<td>bobby ian greg</td>
<td>collins murray murphy</td>
</tr>
<tr>
<td>singh bherusconi sharon</td>
<td>shares referee suggesting</td>
</tr>
<tr>
<td>blair putin abbas</td>
<td>industries testified insisted</td>
</tr>
<tr>
<td>netanyahu brown levy</td>
<td>adding arguing yesterday</td>
</tr>
<tr>
<td>tampa colorado minnesota</td>
<td>bay derby division</td>
</tr>
<tr>
<td>detroit cleveland phoenix</td>
<td>county sox district</td>
</tr>
<tr>
<td>indiana seattle dallas</td>
<td>sole river valley ballet</td>
</tr>
<tr>
<td>policemen helicopters soldiers</td>
<td>remained expressed outst</td>
</tr>
<tr>
<td>suspects demonstrators guards</td>
<td>recommended remains feels</td>
</tr>
<tr>
<td>iraquis personnel detainees</td>
<td>gets resumed sparked</td>
</tr>
<tr>
<td>salt chicken pepper</td>
<td>chicken cream pepper</td>
</tr>
<tr>
<td>chocolate butter cheese</td>
<td>sauce cheese chocolate</td>
</tr>
<tr>
<td>cream sauce bread</td>
<td>salt butter bread</td>
</tr>
</tbody>
</table>

Table 1: Words likely to be generated for singular vector pairs of the LDS transition operator. The operator maps right vectors to left, and the pairs are syntactically and semantically coherent.

This state. Table 1 presents these singular vector pairs. We find they reflect interpretable transition dynamics. In all but the last block, the vectors reflect strict state transitions. However, in the last block contains topical terms about food invariant under \( A \). Overall, we did not find such salient structure in the parameters estimated using SSID.

7.2. POS Tagging

Unsupervised learning of generative discrete state models for text has been shown to capture part-of-speech (POS) information (Christodoulopoulos et al., 2010). In response, we assess the ability of the LDS to also capture POS structure. Token embeddings can be used to predict POS in two ways: (1) by applying a local classifier to each token’s embedding, or (2) by including each token’s embedding as additional features in a lexicalized tagger. For both, we train the tagging model on the Penn Treebank (PTB) train set, which is not included for LDS training. Token embeddings are obtained from Kalman smoothing. We evaluate tagging accuracy on the PTB test set using the 12 ‘universal’ POS tags (Petrov et al., 2011) and the original tags. We contrast the LDS with type embeddings from Word2Vec, trained on the LDS data (Mikolov et al., 2013).

We fit our LDS using a combination of the APNews, New York Times, and RCV1 newswire corpora, about 1B tokens total. We maintain punctuation and casing of the text, but replace all digits with “NUM” and all but the most 200k frequent types with “OOV.” We employ \( r = 4 \) for SSID, \( r = 7 \) for EM, and \( h = 200 \). We add 1000 pseudocounts for each type, by adding \( \frac{1000}{r} \) to each coordinate of \( \mu \).

The LDS hyperparameters were selected by maximizing the accuracy of a local classifier on the PTB dev set. This also included when to terminate EM. For Word2Vec, we performed a broad search over hyperparameters, again maximizing for local POS tagging. Our local classifier was a two-layer neural network with 25 hidden units, which outperformed a linear classifier. The best Word2Vec configuration used the CBOW architecture with a window width of 3. The lexicalized tagger’s hyperparameters were also tuned on the PTB dev set. For the local tagging, we ignored punctuation and few common words types such as “and” in training. Instead, we classified them directly using their majority tag in the training data.

Overall, we found that the LDS and Word2Vec took about 12 hours to train on a single-core CPU. Since the Word2Vec algorithm is simple and the code is heavily optimized, it performs well, but our learning algorithm would have been substantially faster given a larger training set, since the \( \Psi_k \) matrices can be gathered in parallel and the cost of SSID and ASOS is sublinear in the corpus size. In Section 7.4, training the LDS is order of magnitude faster than an RNN.

Our results are shown in Table 2. Left to right, we compare Word2Vec (W2V), SSID, EM initialized with SSID (EM), our baseline lexicalized tagger (LEX), the lexicalized tagger with extra features from LDS token embeddings (LEX + EM), and the lexicalized tagger with type-level Word2Vec embeddings (LEX + SSID).

<table>
<thead>
<tr>
<th>Tags</th>
<th>W2V</th>
<th>SSID</th>
<th>EM</th>
<th>Lex</th>
<th>Lex +EM</th>
<th>Lex +W2V</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>95.00</td>
<td>89.25</td>
<td>96.44</td>
<td>97.97</td>
<td>98.05</td>
<td>98.02</td>
</tr>
<tr>
<td>P</td>
<td>92.58</td>
<td>83.00</td>
<td>94.30</td>
<td>97.28</td>
<td>97.32</td>
<td>97.35</td>
</tr>
</tbody>
</table>

Table 2: POS tagging with universal tags (U) and PTB tags (P).

The first 3 columns perform local classification. First, while SSID is crucial for EM initialization, we found it performed poorly on its own. However, EM outperforms Word2Vec substantially. We expect this is because the LDS explicitly maximizes the likelihood of text sequences, and thus it forces token embeddings to capture the transition dynamics of syntax. All differences are statistically significant at a .05 significance level using the exact binomial test. In Appendix C, we demonstrate the importance of SSID vs. random initialization. The final 3 columns use a carefully-engineered tagger. For universal tags, LDS and Word2Vec both contribute a statistically-significant gain over the baseline (Lex), but their difference is not significant. For PTB tags, we find that Word2Vec achieves a significant gain over LEX, but the LDS does not. We expect that our context-dependent embeddings perform as well as context-independent embeddings since the taggers’ features and test-time inference capture non-local interactions.

7.3. Named Entity Recognition

In Table 3 we consider the effect of unsupervised token features for NER on the Conll 2003 dataset using a lexicalized
We use the popular RNN learning rate schedule where it is constant until performance on held-out data fails to improve, and then it is decreased geometrically until the held-out performance again fails to improve (Mikolov, 2012). We tuned the initial value and decay rate. When initializing with the LDS, small learning rates are crucial: otherwise, the optimization jumps far from where it started.

In Figure 1, we plot perplexity on the dev set vs. the number of training epochs. The time to train the LDS, about 30 minutes, is inconsequential compared to training the RNN (4 days) on a single CPU core. LDS training on the PTB is faster than our experiments above with 1B tokens because we use a small vocabulary and run far fewer EM iterations, in order to prevent overfitting. The RNN baseline converged after 17 training epochs, while using the LDS for initialization allowed it to converge after 12, which amounts to about a day of savings on a single CPU core. Next, in Table 4 we compare the final perplexities on the dev and test sets. We find that initializing with the LDS also provides a better model.

Table 4. Final perplexity for an RNN language model trained using random parameter initialization vs. LDS initialization.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>LDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>130.6</td>
<td>128.1</td>
</tr>
<tr>
<td>test</td>
<td>124.0</td>
<td>122.8</td>
</tr>
</tbody>
</table>

We found that initializing the RNN with LDS parameters trained using SSID, rather than SSID+EM, performed no better than the baseline. Specifically, the best performance was obtained using a high initial learning rate, which allows gradient descent to ignore the SSID values. We expect this is because the method of moments requires lots of data, and the PTB is small. In a setting where one trains the LDS on a very large corpus, it is possible that SSID is effective. Overall, we did not explore initializing the RNN using type-level embeddings such as Word2Vec, since it is unclear how to initialize \( A \) and how to set \( K \) vs. \( C \).

8. Conclusion and Future Work

We have contributed a scalable method for assigning word tokens context-specific low-dimensional representations that capture useful syntactic and semantic structure. Our algorithm requires a single pass over the training data and no painful tuning of learning rates.

Next, we will extend ASOS to new models and improve initialization for alternative RNN architectures, including hierarchical softmaxes, by leveraging not just the LDS parameters, but also the LDS posterior on the training data.

Acknowledgments

This work was partially completed while the first author was an intern at Microsoft Research. We appreciate helpful...
comments from Brendan O’Connor, Ben Marlin, Andrew McCullum, and Qingqing Huang.

References


Van Overschee, Peter and De Moor, B. Subspace identification for linear systems: Theory, implementation. Methods, 1996.
