Compressible Reparametrization of Time-Variant Linear Dynamical Systems

Nico Piatkowski[1](✉) and François Schnitzler[2]

[1] Artificial Intelligence Group, TU Dortmund, 44227 Dortmund, Germany
nico.piatkowski@tu-dortmund.de
[2] Technicolor, 35576 Cesson-sévigné, France
francois.schnitzler.ml@gmail.com

Abstract. Linear dynamical systems (LDS) are applied to model data from various domains—including physics, smart cities, medicine, biology, chemistry and social science—as stochastic dynamic process. Whenever the model dynamics are allowed to change over time, the number of parameters can easily exceed millions. Hence, an estimation of such time-variant dynamics on a relatively small—compared to the number of variables—training sample typically results in dense, overfitted models. Existing regularization techniques are not able to exploit the temporal structure in the model parameters. We investigate a combined reparametrization and regularization approach which is designed to detect redundancies in the dynamics in order to leverage a new level of sparsity. On the basis of ordinary linear dynamical systems, the new model, called ST-LDS, is derived and a proximal parameter optimization procedure is presented. Differences to l_1-regularization-based approaches are discussed and an evaluation on synthetic data is conducted. The results show, that the larger the considered system, the more sparsity can be achieved, compared to plain l_1-regularization.

1 Introduction

Linear dynamical systems (LDS) describe relationships among multiple quantities. The system defines how the quantities evolve over time in response to past or external values. They are important for analyzing multivariate time-series in various domains such as economics, smart-cities, computational biology and computational medicine. This work aims at estimating the transition matrices of finite, time-variant high-dimensional vector time-series.

Large probabilistic models [8,17] are parameterized by millions of variables. Moreover, models of spatio-temporal data like dynamic Bayesian networks (DBN) [3] become large when transition probabilities between time-slices are not time-invariant. This induces problems in terms of tractability and overfitting. A generic solution to these problems is a restriction to *sparse* models. Approaches to find sparse models by penalizing parameter vectors with many non-zero weights are available (e.g., the LASSO [5,15]). However, setting model parameters to zero implies changes to the underlying conditional independence structure [8]. This is not desired if specific relations between variables are to be studied.

S. Michaelis et al. (Eds.): Morik Festschrift, LNAI 9580, pp. 234–250, 2016.
DOI: 10.1007/978-3-319-41706-6_12

To overcome this issue for spatio-temporal data, a combination of reparametrization and regularization has been proposed, called spatio-temporal random fields (STRF) [12], which enables sparse models while keeping the conditional independence structure intact. Although the model in the aforementioned work is presented entirely for discrete data, the underlying concept can be extended to continuous data as well. Here, this idea is investigated and evaluated for multivariate Gaussian data where the conditional independence structure is encoded by the entries of the inverse covariance matrix [8] and a set of transition matrices. It is assumed, that the spatial structure is known and the goal is to find a sparse *representation* of the model's dynamics.

Related Work. In the literature, known approaches that aim at the reduction of model parameters are based on the identification of sparse conditional independence structures which in turn imply sparse parameter vectors. The basic ideas of these approaches can be applied to both, (inverse) covariance matrices and transition matrices. Some important directions are discussed in the following.

General regularization-based methods for sparse estimation may be considered [5, 15], but several approaches for dynamic systems arose in the last decades. In time-varying dynamic Bayesian networks [14], Song et al. describe how to find the conditional independence structure of continuous, spatio-temporal data by performing a kernel reweighting scheme for aggregating observations across time and applying ℓ_1-regularization for sparse structure estimation. In subsequent work, it is shown how to transfer their ideas to spatio-temporal data with discrete domains [7]. The objective function that is used in the latter approach contains a regularization term for the difference of the parameter vectors of consecutive time-slices. Therefore, it is technically the most similar to STRF. However, the estimation is performed locally for each vertex and the resulting local models are heuristically combined to arrive at a global model. It can be shown that this is indeed enough to consistently estimate the neighborhood of each vertex [13].

Statistical properties of conditional independence structure estimation in undirected models are presented in [20]. In particular, the authors investigate (i) the risk consistency and rate of convergence of the covariance matrix and its inverse, (ii) large deviation results for covariance matrices for non-identically distributed observations, and (iii) conditions that guarantee smoothness of the covariances.

Han and Liu [6] present the first analysis of the estimation of transition matrices under a high-dimensional doubly asymptotic framework in which the length and the dimensionality of the time-series are allowed to increase. They provide explicit rates of convergence between the estimator and the population transition matrix under different matrix norms.

ℓ_1-regularization is indeed not the only way for inducing sparsity into the model. Wong et al. [18] show how to incorporate the non-informative Jeffreys hyperprior into the estimation procedure. The main benefits of their approach are the absence of any regularization parameter and approximate unbiasedness of the estimate. However, the resulting posterior function is non-convex and their simulation results indicate that the proposed method tends to underestimate the number of non-zero parameters.

Instead of regularization, score-based methods deliver a combinatorial alternative for structure learning. Therein, multiple independence tests are performed to detect local structures which are finally merged to a global conditional independence structure. Since a large number of tests has to be performed, the approach might not be applicable whenever the number of variables is high. Local search heuristics [10,16] can leverage such complexity issues by restricting the test-space to neighboring structures.

Approaches mentioned so far assume that a specific segmentation of the data in suitable time-slices is already available. Fearnhead [4] developed efficient dynamic programming algorithms for the computation of the posterior over the number and location of changepoints in time-series. Based on this line of research, Xuan and Murphy [19] show how to generalize Fearnheads algorithms to multidimensional time-series. Specifically, they model the conditional independence structure using sparse, ℓ_1-regularized, Gaussian graphical models. The techniques presented therein can be used to identify the maximum a posteriori segmentation of time-series, which is required to apply any of the algorithms mentioned above.

Contribution and Organization. It is shown how to adapt the STRF model [12] to time-variant linear dynamical systems. Two alternatives are discussed, namely a reparametrization of the exponential family form of the system and a reparametrization of the transition matrices. Furthermore, a proximal-algorithm-based optimization procedure [1,11] for the joint estimation of the compressed transition matrices is presented. Finally, we evaluate the proposed procedure on synthetic data in terms of quality and complexity. The results are compared to ℓ_1-regularization and ordinary LDS.

2 Linear Dynamical Systems

Before our spatio-temporal reparametrization can be explained, we introduce time-variant linear dynamical systems and their estimation from data. Let $\boldsymbol{x}_{1:T} := (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ be a n-dimensional real valued time-series. We assume that its autonomous dynamics are fully specified by a finite, discrete-time, affine matrix equation

$$\boldsymbol{x}_t = \boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t \qquad \text{for } 1 < t \leq T \tag{1}$$

with *state* $\boldsymbol{x}_t \in \mathbb{R}^n$, *transition matrix* $\boldsymbol{A}_t \in \mathbb{R}^{n \times n}$ and *noise* $\boldsymbol{\varepsilon}_t \in \mathbb{R}^n$. We call $\boldsymbol{x}_1$ the *initial state* of the system. In total, there are $T-1$ transition matrices $\boldsymbol{A} := (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{T-1})$[1]. Each $\boldsymbol{\varepsilon}_t$ is drawn from the same multivariate Gaussian distribution $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Due to this stochasticity, each $\boldsymbol{x}_t$ with $t > 1$ is a multivariate Gaussian random variable given $\boldsymbol{x}_{t-1}$, with

[1] Notice that $\boldsymbol{A}$ is a short notation for all transition matrices of the system.

$\boldsymbol{x}_t|\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma})$. If the initial state is considered as a random variable too, e.g., $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, the full joint probability density of $\boldsymbol{x}_{1:T}$ may be denoted as:

$$\mathbb{P}_{\boldsymbol{A},\boldsymbol{\Sigma}}(\boldsymbol{x}_{1:T}) = \mathbb{P}_{\boldsymbol{\Sigma}}(\boldsymbol{x}_1) \prod_{t=1}^{T-1} \mathbb{P}_{\boldsymbol{A}_t,\boldsymbol{\Sigma}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \tag{2}$$

If $\boldsymbol{x}_1$ is deterministic instead, one may simply drop the leading factor in (2).

2.1 Parameter Estimation

Estimating the parameters of an LDS is typically done by maximizing the likelihood $\mathcal{L}(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D})$ of a given dataset $\mathcal{D} = \{\boldsymbol{x}^i_{1:T}\}_{i=1}^N$ that contains N realizations of the time-series $\boldsymbol{x}_{1:T}$.

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) = \prod_{i=1}^{N} \mathbb{P}_{\boldsymbol{A},\boldsymbol{\Sigma}}(\boldsymbol{x}^i_{1:T}) \tag{3}$$

Notice that we parameterize the likelihood directly in terms of the inverse covariance matrix. Since non-degenerate covariance matrices are positive definite, such an inverse is guaranteed to exist. Due to numerical convenience, it is common to minimize the average negative log-likelihood $\ell(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) = -\frac{1}{NT} \log \mathcal{L}(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D})$ instead. By plugging (2) into (3) and substituting the Gaussian density for $\mathbb{P}$, the resulting objective function is:

$$\begin{aligned}
\ell(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) &= -\frac{1}{NT} \log \prod_{i=1}^{N} \mathbb{P}_{\boldsymbol{A},\boldsymbol{\Sigma}}(\boldsymbol{x}^i_{1:T}) \\
&= -\frac{1}{NT} \sum_{i=1}^{N} \left(\log \mathbb{P}_{\boldsymbol{\Sigma}}(\boldsymbol{x}^i_1) + \sum_{t=1}^{T-1} \log \mathbb{P}_{\boldsymbol{A},\boldsymbol{\Sigma}}(\boldsymbol{x}^i_{t+1}|\boldsymbol{x}_t) \right) \\
&= C - \frac{1}{2} \log\det\, \boldsymbol{\Sigma}^{-1} + \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{r}_t^{i\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{r}^i_t
\end{aligned} \tag{4}$$

with *residual vector* $\boldsymbol{r}^i_t = \boldsymbol{x}^i_t - \boldsymbol{A}_{t-1}\boldsymbol{x}^i_{t-1}$, constant $C = \frac{1}{2} n \log 2\pi$ and $\top$ indicates the transpose of a vector or matrix. Here, $\mathbb{P}_{\boldsymbol{\Sigma}}(\boldsymbol{x}_1)$ is absorbed into the summation by setting $\boldsymbol{x}_0 := \boldsymbol{0}$ and $\boldsymbol{A}_0 := \boldsymbol{0}$. In the last equation, we made use of the fact that $(\det\, \boldsymbol{\Sigma}^{-1}) = (\det\, \boldsymbol{\Sigma})^{-1}$ since any covariance matrix is positive definite. ℓ is a convex function of the transition matrices and the inverse noise covariance matrix, due to the convexity of $-\log\det\, \boldsymbol{\Sigma}^{-1}$ and $(\boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1})^2$. First or second order optimization procedures may be applied to find the global minimizer of (4) w.r.t. $\boldsymbol{A}$ or $\boldsymbol{\Sigma}$. Hence, it is useful to know the derivatives.

We adopt the notation from [9] whenever an expression involves matrix differential calculus. Let the operator vec : $\mathbb{R}^{m \times n} \to \mathbb{R}^{mn}$ transform a matrix into a vector by stacking the columns of the matrix one underneath the other—vec$(\boldsymbol{M})$

represents the matrix $\boldsymbol{M}$ in column-major order. The partial derivative of ℓ w.r.t. $\boldsymbol{A}_t$ for $1 \leq t < T$ is then

$$\frac{\partial \ell}{\partial \operatorname{vec}(\boldsymbol{A}_t)^\top} = \frac{1}{2NT} \sum_{i=1}^{N} \frac{\partial \left(\boldsymbol{r}_{t+1}^{i\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_{t+1}^{i} \right)}{\partial \operatorname{vec}(\boldsymbol{A}_t)^\top}$$
$$= -\frac{1}{NT} \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} (\boldsymbol{x}_{t+1}^{i} - \boldsymbol{A}_t \boldsymbol{x}_t^{i}) \boldsymbol{x}_t^{i\top} \right)^\top \tag{5}$$

and its partial derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$ is

$$\frac{\partial \ell}{\partial \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^\top} = -\frac{1}{2} \frac{\partial \log \det \boldsymbol{\Sigma}^{-1}}{\partial \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^\top} + \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\partial \left(\boldsymbol{r}_t^{i\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{r}_t^{i} \right)}{\partial \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^\top}$$
$$= -\frac{1}{2} \operatorname{vec}\left(\boldsymbol{\Sigma} + \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{r}_t^{i} \boldsymbol{r}_t^{i\top} \right)^\top$$

Notice that the first order condition $\partial \ell / \partial \operatorname{vec}\left(\boldsymbol{\Sigma}^{-1}\right)^\top = 0$ implies that the minimizer $\boldsymbol{\Sigma}^*$ must be equal to the empirical second moment of the transformed residual vector. A similar closed form can be derived for $\boldsymbol{A}_t^*$ whenever $\sum_{i=1}^{N} \boldsymbol{x}_t^{i} \boldsymbol{x}_t^{i\top}$ is invertible.

2.2 Sparse Estimation

Using closed-form expressions for $\boldsymbol{A}_t^*$ or $\boldsymbol{\Sigma}^{-1*}$ typically results in dense matrices, i.e., solutions with almost no zero entries. This might not be desired, either because sparse solutions allow for faster computation, or because the resulting matrices should reveal insights about the dependency between variables. A way to achieve this is to bias the solution towards sparse matrices by regularizing the objective function:

$$\ell^{\text{reg}}(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) = \ell(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) + g(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1})$$

where g is an arbitrary non-negative function, the *regularizer*, that somehow measures the *complexity* that is induced by $\boldsymbol{A}$ and $\boldsymbol{\Sigma}^{-1}$. Hence, minimizing ℓ^{reg} will produce solutions that trade off quality (in our case: likelihood) against complexity. It is common to choose a norm as regularizer. In particular, the l_1-norm is known to induce sparse solution [5,15]. For the LDS objective (4), this results in

$$\ell^{l_1\text{-LDS}}(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) = \ell(\boldsymbol{A}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) + \lambda \sum_{t=1}^{T-1} \|\boldsymbol{A}_t\|_1 + \delta \|\boldsymbol{\Sigma}^{-1}\|_1 \tag{6}$$

where $\|\cdot\|_1$ is the entry-wise matrix l_1-norm, i.e., $\|\boldsymbol{M}\|_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} |[\boldsymbol{M}]_{i,j}|$ for any $n \times m$ matrix $\boldsymbol{M}$. Here, λ and δ are positive weights which control the strength of the regularization. The larger λ (δ), the smaller will the norm of the resulting $\boldsymbol{A}_t$ ($\boldsymbol{\Sigma}^{-1}$) be. That is, the larger λ or δ, the higher the number of zero entries in $\boldsymbol{A}_t$ or $\boldsymbol{\Sigma}^{-1}$, respectively.

Remark 1. Zeros at the (i,j)-th entry of the inverse covariance matrix correspond to conditional independence between the variables $[\boldsymbol{x}_t]_i$ and $[\boldsymbol{x}_t]_j$[2], given all the other variables $\{[\boldsymbol{x}_t]_k : k \neq i \neq j\}$ [8]. Since $\boldsymbol{\Sigma}^{-1}$ is an inverse covariance matrix, it is symmetric. Hence, it may be interpreted as the (weighted) adjacency matrix of an undirected graphical structure $G(\boldsymbol{\Sigma}^{-1}) = (V, U)$ with $n = |V|$ vertices and an edge set E. If the estimation is carried out via numerical optimization, special care has to be taken to ensure that the estimated $\boldsymbol{\Sigma}^{-1}$ is symmetric and positive definite. Results on the estimation of sparse inverse covariance matrices may be found in [2,5,21]. In what follows, we assume that $\boldsymbol{\Sigma}$ is known. This is in line with the original STRF, where a spatial graphical structure is assumed to be given [12].

Due to the l_1 term, (6) can not be optimized by conventional numerical methods because $|x|$ is not differentiable at $x = 0$. However, if the gradient of (6) is Lipschitz continuous with modulus L, the proximal gradient method is guaranteed to converge with rate $\mathcal{O}(1/k)$ when a fixed stepsize $\eta \in (0, 1/L]$ is used [11].

Recall that we are interested in minimizing (6) w.r.t. all transition matrices $\boldsymbol{A}_t$. Hence, we consider block-wise minimization of the $\boldsymbol{A}_t$. The proximal alternating linearized minimization [1] is a variant of the general proximal gradient algorithm which is designed for a block-wise setting. A closer investigation of (5) shows, that each partial derivative of (6) w.r.t. $\boldsymbol{A}_t$ is indeed Lipschitz continuous. It's block Lipschitz constant is $L_t = \frac{1}{T}\|\boldsymbol{\Sigma}^{-1}\|_F \|\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_t^i \boldsymbol{x}_t^{i\top}\|_F = \|(\partial/\partial\operatorname{vec}(\boldsymbol{A}_t)^\top)(\partial\ell/\partial\operatorname{vec}(\boldsymbol{A}_t)^\top)\|_F$, which is the Frobenius norm of the gradient's Jacobian w.r.t. to $\boldsymbol{A}_t$. This is based on the fact that any differentiable vector-valued function whose gradient has bounded norm is Lipschitz continuous.

Using these moduli of continuity, the optimization consists of iteratively updating all transition matrices. Let $\gamma > 1$. In each iteration, the transition matrices are updated according to

$$\operatorname{vec}\left(\boldsymbol{A}_t^{\text{new}}\right)^\top = \operatorname{prox}_{\gamma L_t}\left(\operatorname{vec}\left(\boldsymbol{A}_t\right)^\top - \frac{1}{\gamma L_t}\frac{\partial \ell}{\partial \operatorname{vec}\left(\boldsymbol{A}_t\right)^\top}\right)$$

with

$$\operatorname{prox}_\lambda^{\|\cdot\|_1}(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}}\left(\|\boldsymbol{y}\|_1 + \frac{\lambda}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right).$$

Moreover, since $\|\cdot\|_1$ is fully separable, it can be shown (see, e.g., [11]) that

$$[\operatorname{prox}_\lambda^{\|\cdot\|_1}(\boldsymbol{x})]_j = \begin{cases} \boldsymbol{x}_j - \lambda, & \boldsymbol{x}_j > \lambda \\ 0, & |\boldsymbol{x}_j| \leq \lambda \\ \boldsymbol{x}_j + \lambda, & \boldsymbol{x}_j < -\lambda \end{cases}.$$

[2] $[\boldsymbol{x}]_i$ represents the i-th component of vector $\boldsymbol{x}$. Moreover, $[\boldsymbol{M}]_{i,j}$ represents the entry in row i and column j of matrix $\boldsymbol{M}$.

2.3 LDS and the Exponential Family

The STRF reparametrization for discrete state Markov random fields is formulated for exponential families [17]. We will now shortly recap the exponential family form of the multivariate Gaussian, which is also known as *information form*. An exponential family with natural parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ may be denoted as

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - B(\boldsymbol{\theta})\right) \tag{7}$$

with log partition function $B(\boldsymbol{\theta}) = \log \int \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle\right) \mathrm{d}\boldsymbol{x}$[3]. In case of the multivariate Gaussian, parameter and sufficient statistic $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^d$ are given by

$$\boldsymbol{\theta} = \begin{pmatrix} -\frac{1}{2}\operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \end{pmatrix} \text{ and } \phi(\boldsymbol{x}) = \begin{pmatrix} \operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^\top) \\ \boldsymbol{x} \end{pmatrix},$$

respectively. Moreover, the closed form of $B(\boldsymbol{\theta})$ can be computed by the n-dimensional Gaussian integral:

$$\begin{aligned} B(\boldsymbol{\theta}) &= \log \int \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle\right) \mathrm{d}\boldsymbol{x} \\ &= \log \int \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} \;+\; \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \mathrm{d}\boldsymbol{x} \\ &= \log \left(\sqrt{(2\pi)^n \det\, \boldsymbol{\Sigma}^{-1}} \exp\left(\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right). \end{aligned}$$

Plugging this into Eq. (7) and rearranging, one arrives at the well known expression for the multivariate Gaussian density:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - B(\boldsymbol{\theta})\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \det\, \boldsymbol{\Sigma}^{-1}}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} \;+\; \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \;-\; \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n \det\, \boldsymbol{\Sigma}^{-1}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \end{aligned}$$

Based on this equivalence, the joint density (2) of an LDS can also be rewritten in terms of exponential families (7)

$$\begin{aligned} \mathbb{P}_{\boldsymbol{A},\boldsymbol{\Sigma}}(\boldsymbol{x}_{1:T}) &= \mathbb{P}_{\boldsymbol{\Sigma}}(\boldsymbol{x}_1) \prod_{t=1}^{T-1} \mathbb{P}_{\boldsymbol{A}_t,\boldsymbol{\Sigma}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \\ &= \mathbb{P}_{\boldsymbol{\theta}_1}(\boldsymbol{x}_1) \prod_{t=1}^{T-1} \mathbb{P}_{\boldsymbol{\theta}_{t+1}}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \\ &= \exp\left(\sum_{t=1}^{T} \langle \boldsymbol{\theta}_t, \phi_t(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})\rangle - B(\boldsymbol{\theta}_t, \boldsymbol{x}_{t-1})\right) \end{aligned}$$

[3] The log partition function is usually denoted by $A(\boldsymbol{\theta})$. Since the symbol A is already reserved for transition matrices, we denote the log partition function with B instead.

where we used the exponential family form of each $\boldsymbol{x}_t|\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma})$ and hence, the parameters and sufficient statistic are

$$\boldsymbol{\theta}_t = \begin{pmatrix} -\frac{1}{2}\operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \\ \operatorname{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{t-1}) \end{pmatrix}, \quad \phi_t(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \begin{pmatrix} \operatorname{vec}(\boldsymbol{x}_t\boldsymbol{x}_t^\top) \\ \operatorname{vec}(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^\top) \end{pmatrix}.$$

Again, we set $\boldsymbol{x}_0 := \boldsymbol{0}$ and $\boldsymbol{A}_0 := \boldsymbol{0}$ to compactify notation. To remove the functional dependence between the local log-partition functions $B(\boldsymbol{\theta}_t, \boldsymbol{x}_{t-1})$ and $\boldsymbol{x}_{t-1}$, we include the corresponding term $-\frac{1}{2}\boldsymbol{x}_{t-1}^\top\boldsymbol{A}_{t-1}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1}$ directly into the parameters and sufficient statistics:

$$\tilde{\boldsymbol{\theta}}_t = \begin{pmatrix} -\frac{1}{2}\operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \\ \operatorname{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{t-1}) \\ -\frac{1}{2}\operatorname{vec}(\boldsymbol{A}_{t-1}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{t-1}) \end{pmatrix}, \quad \tilde{\phi}_t(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \begin{pmatrix} \operatorname{vec}(\boldsymbol{x}_t\boldsymbol{x}_t^\top) \\ \operatorname{vec}(\boldsymbol{x}_t\boldsymbol{x}_{t-1}^\top) \\ \operatorname{vec}(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}^\top) \end{pmatrix}.$$

Finally, the joint probability of the LDS in exponential family form is

$$\mathbb{P}_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x}_{1:T}) = \exp\left(\langle\tilde{\boldsymbol{\theta}}, \tilde{\phi}(\boldsymbol{x}_{1:T})\rangle - B(\tilde{\boldsymbol{\theta}})\right)$$

where, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_T)^\top$, $\tilde{\phi} = (\tilde{\phi}_1(\boldsymbol{x}_1, \boldsymbol{x}_0), \tilde{\phi}_2(\boldsymbol{x}_2, \boldsymbol{x}_1), \ldots, \tilde{\phi}_T(\boldsymbol{x}_T, \boldsymbol{x}_{T-1}))^\top$, and $B(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^T B(\tilde{\boldsymbol{\theta}}_t)$ are the corresponding parameter, sufficient statistics and log partition function, respectively.

This representation has several drawbacks when compared to the native representation in terms of transition matrices. An obvious disadvantage is, that multiple copies of $\boldsymbol{\Sigma}^{-1}$ are encoded into the parameters. Moreover, the transition matrices can only be recovered via inversion of $\boldsymbol{\Sigma}^{-1}$ and subsequent matrix multiplication with the lower part of $\boldsymbol{\theta}_t$ which encodes $\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_{t-1}$. Hence, $\mathcal{O}(n^3)$ flops are required to extract $\boldsymbol{A}_{t-1}$ from $\boldsymbol{\theta}_t$ which might be prohibitive in a large system.

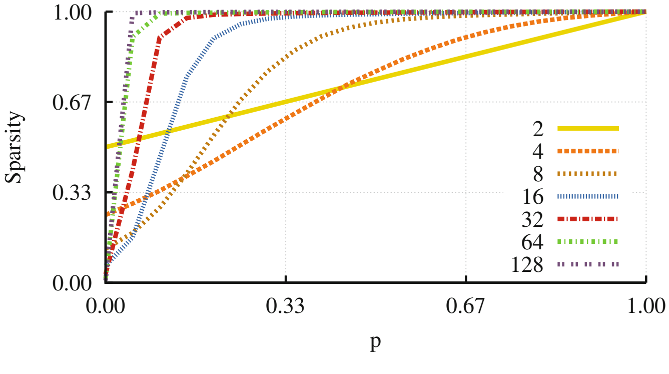


Fig. 1. X-axis: Parameter p of the Bernoulli distribution of the entries of the lower triangular matrix $\tilde{\boldsymbol{L}}$. Y-axis: Average sparsity of $(\tilde{\boldsymbol{L}} + 10\boldsymbol{I}_n)(\tilde{\boldsymbol{L}} + 10\boldsymbol{I}_n)^\top$.

3 Reparametrization of LDS

The main goal of this work is a sparse reparametrization of LDS that does not alter the dependences which are encoded in the transition matrices. If l_1-regularization is applied to a transition matrix $\boldsymbol{A}_t$, some of its entries will be pushed to 0, and hence, some flow of information between variables is prohibited. Moreover, if a particular value of $\boldsymbol{A}_t$ does not change much over time, i.e., $[\boldsymbol{A}_t]_{i,j} \approx c$ for all $1 \leq t < T$, l_1-regularization can not exploit this redundancy. Here, we aim at finding an alternative representation that is able to sparsify such redundancies while keeping small interactions between variables intact. For discrete state Markov random fields, this task has already been solved by STRF. The core of STRF is a spatio-temporal reparametrization of the exponential family

$$\boldsymbol{\theta}_t(\boldsymbol{\Delta}) = \sum_{i=1}^{t} \frac{1}{t-i+1} \boldsymbol{\Delta}_i$$

with l_1 and l_2 regularization of the $\boldsymbol{\Delta}_i$.

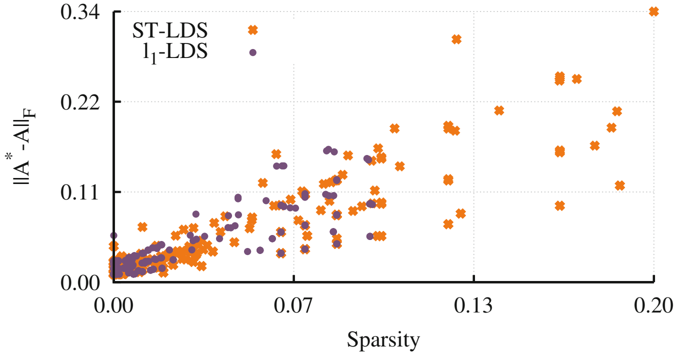


Fig. 2. Sparsity vs. Error. X-axis: Sparsity of estimated transition matrices. Y-axis: Estimation error of transition matrices, measured in Frobenius norm $\|\boldsymbol{A}^* - \boldsymbol{A}\|_F$.

As already mention at the end of Sect. 2, extracting the transition matrices from the exponential family form of an LDS is rather expensive. In practical applications of LDS, the transition matrices are of special interest. Either because a prediction of future states of the system has to be computed, or if particular interactions between variables are investigated. Therefore, we dismiss the exponential family representation and perform the reparametrization w.r.t. the transition matrices.

$$\boldsymbol{A}_t(\boldsymbol{\Delta}) = \sum_{i=1}^{t} \frac{1}{t-i+1} \boldsymbol{\Delta}_i$$

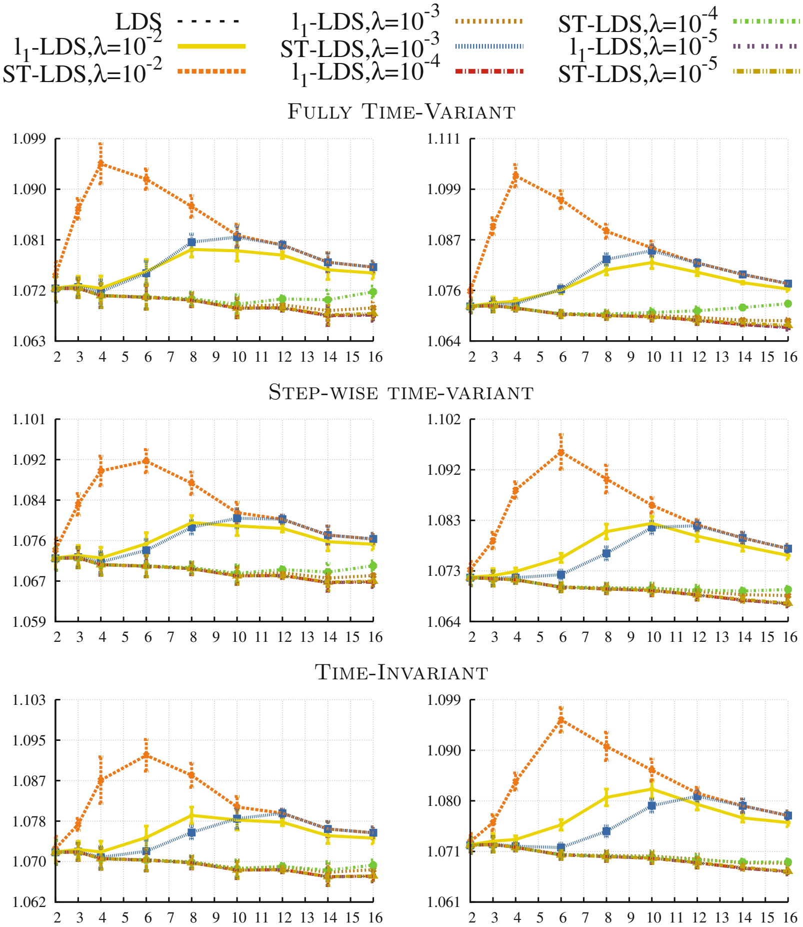


Fig. 3. X-axis: number of variables n. Y-axis: Normalized negative log-likelihood ℓ/n. Left: $T = 4$; Right: $T = 8$. Lower is better.

Analogous to (6), this results in the objective function

$$\ell^{\text{ST-LDS}}(\boldsymbol{\Delta}, \boldsymbol{\Sigma}^{-1}, \mathcal{D}) = \ell(\boldsymbol{A}(\boldsymbol{\Delta}), \boldsymbol{\Sigma}^{-1}, \mathcal{D}) + \lambda \sum_{t=1}^{T-1} \|\boldsymbol{\Delta}_t\|_1 \tag{8}$$

with $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \ldots, \boldsymbol{\Delta}_{T-1})$. Notice that we perform only l_1-regularization of $\boldsymbol{\Delta}$, since the results in [12] suggest that the impact of l_2-regularization on the sparse reparametrization is neglectable. In addition, $\boldsymbol{\Sigma}^{-1}$ is treated as a constant as explained in Remark 1.

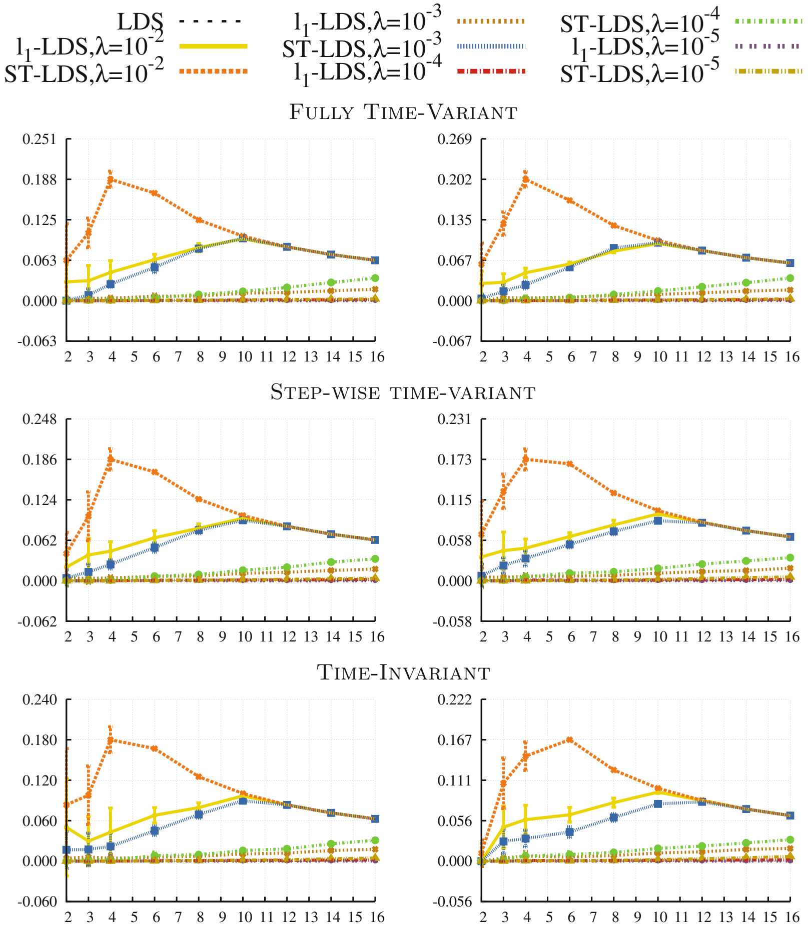


Fig. 4. X-axis: number of variables n. Y-axis: Sparsity of estimated transition matrices. Left: $T = 4$; Right: $T = 8$. Higher is better.

The partial derivatives of ℓ w.r.t. $\boldsymbol{\Delta}_t$ are required to apply the proximal algorithm from Sect. 2.2. We apply the matrix chain rule (see, e.g., [9]) to get

$$\frac{\partial \ell}{\partial \operatorname{vec}(\boldsymbol{\Delta}_t)^\top} = \left(\frac{\partial \ell}{\partial \operatorname{vec}(\boldsymbol{A}(\boldsymbol{\Delta}))^\top}\right)\left(\frac{\partial \operatorname{vec}(\boldsymbol{A}(\boldsymbol{\Delta}))}{\partial \operatorname{vec}(\boldsymbol{\Delta}_t)^\top}\right)$$

with

$$\frac{\partial [\boldsymbol{A}_{t'}(\boldsymbol{\Delta})]_{l,r}}{\partial [\boldsymbol{\Delta}_t]_{i,j}} = \begin{cases} \frac{1}{t'-t+1}, & t' \geq t \wedge i = l \wedge j = r \\ 0, & \text{else} \end{cases} .$$

The block Lipschitz constant $U_t = \sqrt{\sum_{t'=t}^{T-1} \left(n/(t'-t+1)\right)^2}$ of $\boldsymbol{A}(\boldsymbol{\Delta})$ w.r.t. $\boldsymbol{\Delta}_t$ is derived as described in Sect. 2.2, i.e.,

$$U_t = \left\| \left(\frac{\partial}{\partial \operatorname{vec}\left(\boldsymbol{\Delta}_t\right)^\top} \right) \left(\frac{\partial \ell}{\partial \operatorname{vec}\left(\boldsymbol{\Delta}_t\right)^\top} \right) \right\|_F .$$

Now, since $f = \ell \circ \boldsymbol{A}$ is the composition of two Lipschitz continuous functions, $U_t L_t$ is the t-th block Lipschitz constant of $f(\boldsymbol{\Delta}) = \ell(\boldsymbol{A}(\boldsymbol{\Delta}), \boldsymbol{\Sigma}^{-1}, \mathcal{D})$.

4 Experiments

Experiments are conducted in order to investigate and compare the (i) loss, (ii) sparsity and (iii) estimated transition matrices of the following methods:

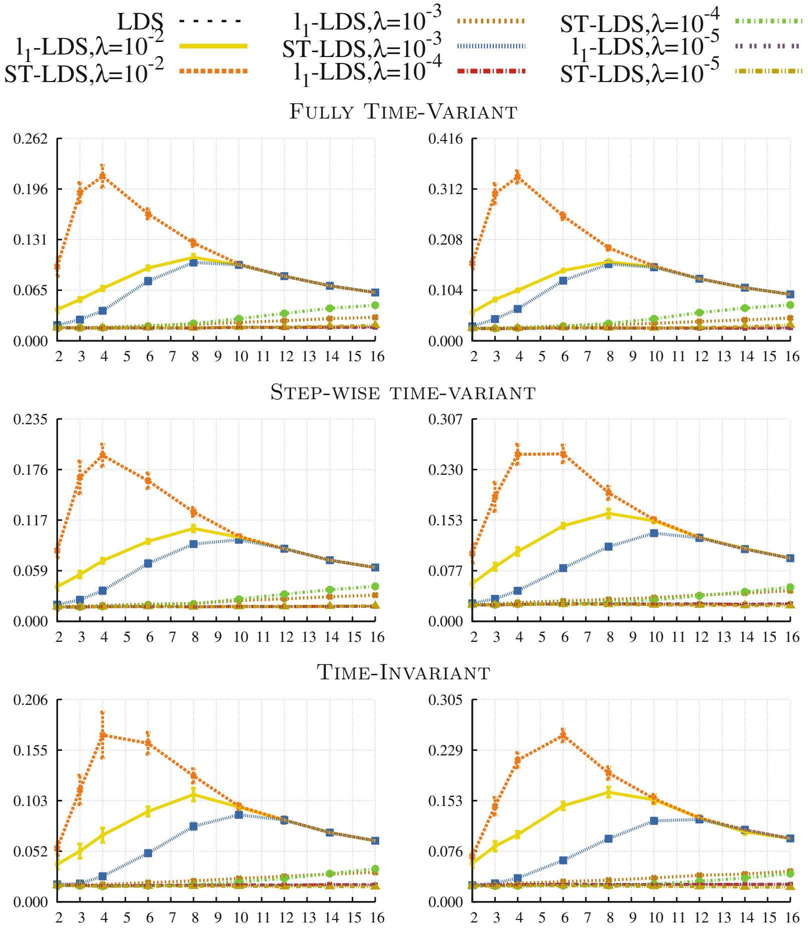
- Plain time-variant LDS as defined in (1) with objective function (4)
- l_1-LDS with objective function (6)
- ST-LDS with objective function (8)

Here, *sparsity* is defined as the fraction of zero-entries in a parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, i.e., $\operatorname{sparsity}(\boldsymbol{\theta}) = \frac{1}{d}\sum_{i=1}^{d} \mathbb{1}(\boldsymbol{\theta}_i = 0)$. The indicator function $\mathbb{1}(\text{expr})$ evaluates to 1 iff expr is true.

The synthetic data for the experimental evaluation is generated by the following stochastic process:

1. Fix the number of variables n, time-steps T and samples N.
2. Generate a random inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$. This is done by generating a lower triangular binary matrix $\tilde{\boldsymbol{L}}$ where each entry is draw independently from a Bernoulli distribution with parameter p. The sign of each non-zero off-diagonal entry is determined by drawing from another Bernoulli with parameter $1/2$. Then, the $n \times n$ up-scaled identity matrix $10\boldsymbol{I}_n$ is added to $\tilde{\boldsymbol{L}}$ and the result is multiplied by its own transpose, i.e., $\tilde{\boldsymbol{\Sigma}}^{-1} = (\tilde{\boldsymbol{L}} + 10\boldsymbol{I}_n)(\tilde{\boldsymbol{L}} + 10\boldsymbol{I}_n)^\top$. The implied $\tilde{\boldsymbol{\Sigma}}$ is normalized in order to have unit variances. Figure 1 shows the sparsity of the final inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ as a function of n and p.
3. Generate $T-1$ random transition matrices $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{t-1}$. The entries are drawn independently from a uniform distribution over $[-\omega, \omega]$:
   (a) For $[\boldsymbol{A}_t]_{i,j}$ and all $1 \leq t < T$. (Fully Time-Variant)
   (b) For $[\boldsymbol{A}_1]_{i,j}$ and $[\boldsymbol{A}_{T/2}]_{i,j}$ and then copied to all $[\boldsymbol{A}_t]_{i,j}$ with $1 < t < T/2$ and $T/2 < t < T$, respectively. (Step-wise time-variant)
   (c) For $[\boldsymbol{A}_1]_{i,j}$ and then copied to all $[\boldsymbol{A}_t]_{i,j}$ for all $1 < t < T$. (Time-Invariant)
4. For $i = 1$ to N
   (a) Draw $\boldsymbol{x}_1^i$ from $\mathcal{N}(0, \boldsymbol{\Sigma})$.
   (b) For $t = 2$ to T
      i. Draw $\boldsymbol{\varepsilon}_t$ from $\mathcal{N}(0, \boldsymbol{\Sigma})$.
      ii. Compute $\boldsymbol{x}_t^i = \boldsymbol{A}_{t-1}\boldsymbol{x}_{t-1}^i + \boldsymbol{\varepsilon}_t$.

This procedure is applied for $n \in \{2, 4, 6, 8, 10, 12, 14, 16\}$, $T \in \{4, 8\}$ and $N = 10000$. Random covariance matrices are generated with $p = 1/4$ and random transition matrices are generated with $\omega = 1/n$. For each combination of n and T, 10 datasets are sampled, which makes a total of 1.6×10^6 data points. The evaluation of regularized methods l_1-LDS and ST-LDS is carried out with $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. All models are estimated by the proximal algorithm, described in Sect. 2.2. In case of an unregularized objective, the proximal algorithm reverts to block-wise gradient descent.



Fully Time-Variant

Step-wise time-variant

Time-Invariant

Fig. 5. X-axis: number of variables n. Y-axis: Estimation error of transition matrices, measured in Frobenius norm $\|\boldsymbol{A}^* - \boldsymbol{A}\|_F$. Left: $T = 4$; Right: $T = 8$. Lower is better.

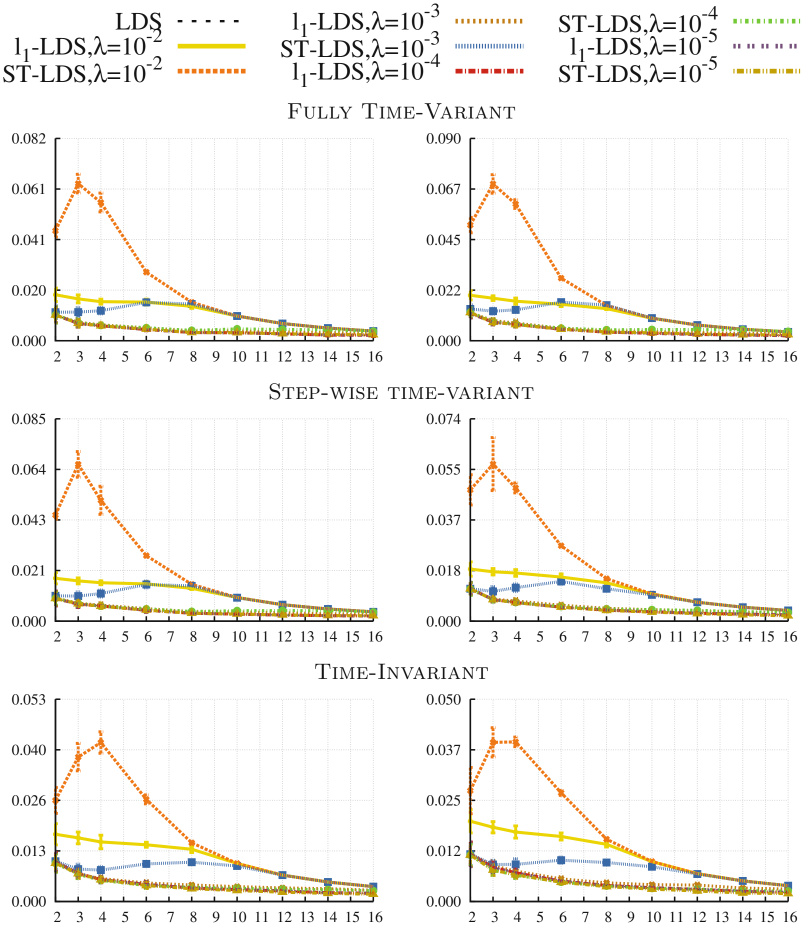


Fig. 6. X-axis: number of variables n. Y-axis: Estimation error of transition matrices, measured in maximum norm $\|\boldsymbol{A}^* - \boldsymbol{A}\|_\infty$. Left: $T = 4$; Right: $T = 8$. Lower is better.

4.1 Likelihood and Sparsity

Results for the average negative log-likelihood (4) and sparsity of the corresponding transition matrices are depicted in Figs. 3 and 4, where each point is averaged over 10 random data sets. Comparing results among different model sizes requires normalization of the loss function values by the corresponding number of variables, hence, Fig. 3 shows ℓ/n. Plots on the left contain results for $T = 4$ time-steps and plots on the right results for $T = 8$ time-steps, respectively. In all cases, a larger value of λ corresponds to more sparsity and a larger loss. Note, however, that the regularization parameter has a different impact on l_1-LDS and ST-LDS models, i.e., sparsity and loss of ST-LDS models with $\lambda = 10^k$ are in the range of l_1-LDS models with $\lambda = 10^{k-1}$. The results suggest the existence of a phase transition which is clearly visible for ST-LDS with $\lambda = 10^{-2}$ and l_1-LDS $\lambda = 10^{-3}$: for small (in terms of n) models, the loss of l_1-LDS is larger than the loss of ST-LDS. However, it can be seen in Fig. 3 that there exists n_0 from which on this relation is interchanged, i.e., the loss of l_1-LDS is lower than the loss of ST-LDS. Remarkably, the sparsity plots (Fig. 4) of the corresponding transition matrices show a similar behavior. Starting from the same n_0, the sparsity of ST-LDS with $\lambda = 10^{-2}$ and the sparsity of l_1-LDS with $\lambda = 10^{-3}$ converge. The point of the phase transition and it's strength depend on the number of time-steps and the type of transition matrices. In case of ST-LDS models with $\lambda = 10^{-5}$ however, the loss is close to that of plain LDS model and the sparsity is larger than that of the corresponding l_1-LDS models. Moreover, the sparsity increases with an increasing number of variables.

4.2 Estimation Error and Sparsity

For each random dataset, we store the original transition matrices $\boldsymbol{A}^*$. This allows us, to investigate the estimation error in terms of the Frobenius norm $\|\boldsymbol{A}^* - \boldsymbol{A}\|_F$ and maximum norm $\|\boldsymbol{A}^* - \boldsymbol{A}\|_\infty$, as shown in Figs. 5 and 6. Again, each point is averaged over 10 random data sets. The ranking of the methods in terms of estimation error is coherent with the sparsity results. While the Frobenius-norm-error increases with an increasing number of variables, the maximum-norm-error is almost zero for all methods with $\lambda \leq 10^{-3}$. While the maximum-norm-error of ST-LDS with $\lambda = 10^{-4}$ is close to 0, the sparsity of the corresponding model increases with an increasing number of variables. Moreover, it's sparsity is higher than the sparsity of the corresponding l_1-LDS model. Finally, the trade-off between sparsity and estimation error is depicted in Fig. 2. Each error-sparsity pair represents one run of the corresponding method. Transition matrices which are estimated with the plain LDS model are completely dense in any case. In general, ST-LDS is able to produce models with a higher sparsity while incorporating a larger error. Notice, however, that some ST-LDS models achieve about twice the sparsity as l_1-LDS models but with the same (rather low) estimation error.

5 Conclusion

In this article, we investigated a combined reparametrization and regularization approach which is designed to detect redundancies in the dynamics of linear dynamical systems. Based on ordinary linear dynamical systems, the new model, called ST-LDS, was derived and a proximal parameter optimization procedure was presented. Expensive line-search techniques or similar step-size adaption techniques were avoided by deriving the block Lipschitz constants of the corresponding objective function w.r.t. the new reparametrization. Differences to l_1-regularization-based approaches were discussed and an evaluation on synthetic data was carried out. The results show, that with an increasing size of an ST-LDS, the estimation error is close to that of an ordinary LDS while achieving more sparsity than l_1-regularization-based models. An investigation of spatio-temporal regression models with non-Gaussian noise is an appealing direction for future research, since many real world phenomena might be explained better by other probability distributions.

Acknowledgement. This work has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Data Analysis”, project A1.

References

1. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
2. Cai, T., Liu, W., Luo, X.: A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J. Am. Stat. Assoc. **106**(494), 594–607 (2011)
3. Dagum, P., Galper, A., Horvitz, E.: Dynamic network models for forecasting. In: Proceedings of the 8th Annual Conference on Uncertainty in Artificial Intelligence, pp. 41–48 (1992)
4. Fearnhead, P.: Exact Bayesian curve fitting and signal segmentation. IEEE Trans. Signal Process. **53**(6), 2160–2166 (2005)
5. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)
6. Han, F., Liu, H.: Transition matrix estimation in high dimensional time series. In: Proceedings of the 30th International Conference on Machine Learning, pp. 172–180 (2013)
7. Kolar, M., Song, L., Ahmed, A., Xing, E.P.: Estimating time-varying networks. Ann. Appl. Stat. **4**(1), 94–123 (2010)
8. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
9. Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd edn. Wiley, Chichester (1999)
10. Rodrigues de Morais, S., Aussem, A.: A novel scalable and data efficient feature subset selection algorithm. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 298–312. Springer, Heidelberg (2008)
11. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Opt. **1**(3), 127–239 (2014)
12. Piatkowski, N., Lee, S., Morik, K.: Spatio-temporal random fields: compressible representation and distributed estimation. Mach. Learn. **93**(1), 115–139 (2013)
13. Ravikumar, P., Wainwright, M.J., Lafferty, J.D.: High-dimensional ising model selection using ℓ_1-regularized logistic regression. Ann. Appl. Stat. **38**(3), 1287–1319 (2010)
14. Song, L., Kolar, M., Xing, E.P.: Time-varying dynamic Bayesian networks. Adv. Neural Inf. Process. Syst. **22**, 1732–1740 (2009)
15. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Royal Stat. Soc. Ser. B **67**(1), 91–108 (2005)
16. Trabelsi, G., Leray, P., Ben Ayed, M., Alimi, A.M.: Dynamic MMHC: a local search algorithm for dynamic bayesian network structure learning. In: Tucker, A., Höppner, F., Siebes, A., Swift, S. (eds.) IDA 2013. LNCS, vol. 8207, pp. 392–403. Springer, Heidelberg (2013)

17. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)
18. Wong, E., Awate, S., Fletcher, T.: Adaptive sparsity in gaussian graphical models. JMLR W&CP **28**, 311–319 (2013)
19. Xuan, X., Murphy, K.: Modeling changing dependency structure in multivariate time series. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1055–1062. ACM (2007)
20. Zhou, S., Lafferty, J.D., Wasserman, L.A.: Time varying undirected graphs. Mach. Learn. **80**(2–3), 295–319 (2010)
21. Zhou, S., Rütimann, P., Xu, M., Bühlmann, P.: High-dimensional covariance estimation based on gaussian graphical models. J. Mach. Learn. Res. **12**, 2975–3026 (2011)