Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Shaojie Chen , hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature: Shaojie Chen Date: 06/19/2016

List any pre-prints: arXiv

Relevant Conference publication(s) (submitted, accepted, or published): NA

Justification for re-publication: NA

Research Highlights (Required)

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

-
-
-
-
-

Pattern Recognition Letters
journal homepage: www.elsevier.com

An M-estimator for reduced-rank system identification

Shaojie Chen[a,**], Kai Liu[b], Yuguang Yang[c], Yuting Xu[a], Seonjoo Lee[d], Martin Lindquist[a], Brian Caffo[a], Joshua T. Vogelstein[e,f]

[a]*Dept. of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA*
[b]*Dept. of Neuroscience, Johns Hopkins University, USA*
[c]*Dept. of Chemical and Biomolecular Engineering, Johns Hopkins University, USA*
[d]*Dept. of Psychiatry and Department of Biostatistics, Columbia University, USA*
[e]*Child Mind Institute, USA*
[f]*Dept. of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, USA*

ABSTRACT

High-dimensional time-series data from a wide variety of domains, such as neuroscience, are being generated every day. Fitting statistical models to such data, to enable parameter estimation and time-series prediction, is an important computational primitive. Existing methods, however, are unable to cope with the high-dimensional nature of these data, due to both computational and statistical reasons. We mitigate both kinds of issues by proposing an M-estimator for Reduced-rank System IDentification (MR. SID). A combination of low-rank approximations, ℓ_1 and ℓ_2 penalties, and some numerical linear algebra tricks, yields an estimator that is computationally efficient and numerically stable. Simulations and real data examples demonstrate the usefulness of this approach in a variety of problems. In particular, we demonstrate that MR. SID can accurately estimate spatial filters, connectivity graphs, and time-courses from native resolution functional magnetic resonance imaging data. MR. SID therefore enables big time-series data to be analyzed using standard methods, readying the field for further generalizations including nonlinear and non-Gaussian state-space models.

1. Introduction

High-dimensional time-series data are becoming increasingly abundant across a wide variety of domains, spanning economics Johansen (1988), neuroscience (Friston et al., 2003), and cosmology (Xie et al., 2013). Fitting statistical models to such data, to enable parameter estimation and time-series prediction, is an important computational primitive. Linear dynamical system (LDS) models are amongst the most popular and powerful, because of their intuitive nature and ease of implementation (Kalman, 1963). The famous Kalman Filter-Smoother is one of the most popular and powerful tools for time-series prediction with an LDS, given known parameters (Kalman, 1960). In practice, however, for many LDS's, the parameters are unknown and must be estimated in a process often called *system identification* (Ljung, 1998). To the best of our knowledge, currently there does not exist a methodology that provides parameter estimates and predictions from ultra-high-dimensional time-series data (e.g. $p > 10{,}000$).

The challenges associated with high-dimensional time-series estimation and prediction are multifold. First, naïvely, such models include dense $p \times p$ matrices, which are often too large to store, much less invert in memory. Several recent efforts to invert large sparse matrices using a series of computational tricks show promise, though they are still extremely computationally expensive (Hsieh et al., 2013; Banerjee et al., 2013). Second, estimators behave poorly due to numerical instability. Reduced-rank LDS models can partially address this problem by reducing the number of latent states. (CHEN et al., 1989). However, without further constraints, the dimensionality of the latent states would be reduced to such an extent that it would significantly decrease the predictive capacity of the resulting model. Third, even after addressing these problems, the time to compute all the necessary quantities can be overly burdensome. Distributed memory implementations, such as those built with Spark, might help overcome this problem. However, it would lead to additional costs and set-up burden, as it would require a

**Corresponding author: Tel.: +1-410-409-8132;
e-mail: `pzcsj76@gmail.com` (Shaojie Chen)

Spark cluster (Zaharia et al., 2010).

We address all three of these issues with our M-estimator for Reduced-rank System IDentification (Mr. Sid). By assuming the dimensionality of the latent state space is small (i.e. reduced-rank), relative to the observed space dimensionality, we can significantly improve computational tractability and estimation accuracy. By further penalizing the estimators, with ℓ_1 and/or ℓ_2 penalties, via utilizing prior knowledge on the structure of the parameters, we gain further estimation accuracy in this high-dimensional but relatively low-sample size regime. Finally, by employing several numerical linear algebra tricks, we can reduce the computational burden significantly.

These three techniques combined enable us to obtain highly accurate estimates in a variety of simulation settings. Mr. Sid is, in fact, a generalization of the now classic Baum-Welch expectation maximization algorithm, commonly used for system identification in much lower dimensional linear dynamical systems (Rabiner, 1989). We show numerically that the hyperparameters can be selected to minimize prediction error on held-out data. Finally, we use Mr. Sid to estimate functional connectomes from the motor cortex. Mr. Sid enables us to estimate the regions, rather than imposing some prior parcellation on the data, as well as estimate sparse connectivity between regions. Mr. Sid reliably estimates these connectomes, as well as predicts the held-out time-series data. To our knowledge, this is the first time a single unified approach has been used to estimate partitions and functional connectomes directly from the high-dimensional data.

This work presents a new analysis of a model which has only been implemented in low-dimensional settings, and paves the way for high-dimensional implementation. Though primitive, it is a first step for essentially any high-dimensional time series analysis, control system identification, and spatiotemporal analysis. To enable extensions, generalizations, and additional applications, the code for the core functions and generating each of the figures is freely available on Github (`https://github.com/shachen/PLDS/`).

2. The first page

Avoid using abbreviations in the title. Next, list all authors with their first names or initials and surnames (in that order). Indicate the author for correspondence (see elsarticle documentation).

Present addresses can be inserted as footnotes. After having listed all authors' names, you should list their respective affiliations. Link authors and affiliations using superscript lower case letters.

2.1. The Abstract

An Abstract is required for every paper; it should succinctly summarize the reason for the work, the main findings, and the conclusions of the study. The abstract should be no longer than 200 words. Do not include artwork, tables, elaborate equations or references to other parts of the paper or to the reference listing at the end. "Comment" papers are exceptions, where the commented paper should be referenced in full in the Abstract.

The reason is that the Abstract should be understandable in itself to be suitable for storage in textual information retrieval systems.

Example of an abstract: A biometric sample collected in an uncontrolled outdoor environment varies significantly from its indoor version. Sample variations due to outdoor environmental conditions degrade the performance of biometric systems that otherwise perform well with indoor samples. In this study, we quantitatively evaluate such performance degradation in the case of a face and a voice biometric system. We also investigate how elementary combination schemes involving min-max or z normalization followed by the sum or max fusion rule can improve performance of the multi-biometric system. We use commercial biometric systems to collect face and voice samples from the same subjects in an environment that closely mimics the operational scenario. This realistic evaluation on a dataset of 116 subjects shows that the system performance degrades in outdoor scenarios but by multimodal score fusion the performance is enhanced by 20%. We also find that max rule fusion performs better than sum rule fusion on this dataset. More interestingly, we see that by using multiple samples of the same biometric modality, the performance of a unimodal system can approach that of a multimodal system.

3. The main text

Please divide your article into (numbered) sections (You can find the information about the sections at `http://www.elsevier.com/wps/find/journaldescription.cws_home/505619/authorinstructions`). Ensure that all tables, figures and schemes are cited in the text in numerical order. Trade names should have an initial capital letter, and trademark protection should be acknowledged in the standard fashion, using the superscripted characters for trademarks and registered trademarks respectively. All measurements and data should be given in SI units where possible, or other internationally accepted units. Abbreviations should be used consistently throughout the text, and all nonstandard abbreviations should be defined on first usage (**?**).

3.1. Tables, figures and schemes

Graphics and tables may be positioned as they should appear in the final manuscript. Figures, Schemes, and Tables should be numbered. Structures in schemes should also be numbered consecutively, for ease of discussion and reference in the text. **Figures should be maximum half a page size.** All numbers and letters in figures and diagrams should be at least of the same font size as that of the figure caption.

Depending on the amount of detail, you can choose to display artwork in one column (20 pica wide) or across the page (42 pica wide). Scale your artwork in your graphics program before incorporating it in your text. If the artwork turns out to be too large or too small, resize it again in your graphics program and re-import it. The text should not run along the sides of any figure. This is an example for citation (**?**).

You might find positioning your artwork within the text difficult anyway. In that case you may choose to place all artwork at

Table 1. Summary of different works pertaining to face and speech fusion

Study	Algorithm used	DB Size	Covariates of interest	Top individual performance	Fusion Performance
UK-BWG (Mansfield et al., 2001)	Face, voice: Commercial	200	Time: 1–2 month separation (indoor)	TAR* at 1% FAR# Face: 96.5% Voice: 96%	–
Brunelli (Brunelli and Falavigna, 1995)	Face: Hierarchical correlation Voice: MFCC	87	Time: 3 sessions, time unknown (indoor)	Face: TAR = 92% at 4.5% FAR Voice: TAR = 63% at 15% FAR	TAR =98.5% at 0.5% FAR
Jain (Jain et al., 1999)	Face: Eigenface Voice: Cepstrum Coeff. Based	50	Time: Two weeks (indoor)	TAR at 1% FAR Face: 43% Voice: 96.5% Fingerprint: 96%	Face + Voice + Fingerprint = 98.5%
Sanderson (Sanderson and Paliwal, 2002)	Face: PCA Voice: MFCC	43	Time: 3 sessions (indoor) Noise addition to voice	Equal Error Rate Face: 10% Voice: 12.41%	Equal Error Rate 2.86%
Proposed study	Face, voice: Commercial	116	Location: Indoor and Outdoor (same day) Noise addition to eye coordinates	TARs at 1% FAR Indoor-Outdoor Face: 80% Voice: 67.5%	TAR = 98% at 1% FAR

*TAR–True Acceptance Rate # FAR–False Acceptance Rate

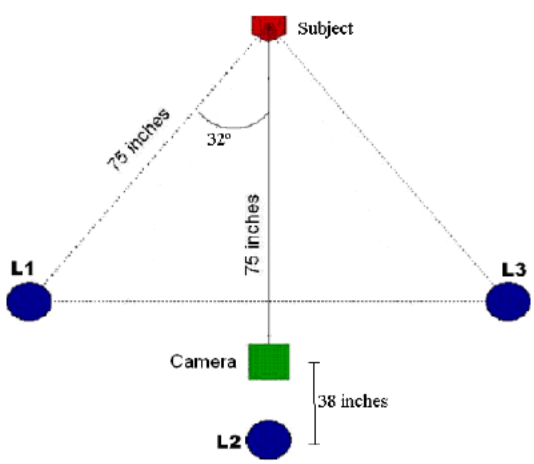

Fig. 1. Studio setup for capturing face images indoor. Three light sources L1, L2, L3 were used in conjunction with normal office lights.

the end of the text and insert a marker in the text at the desired place. In any case, please keep in mind that the placement of artwork may vary somewhat in relation to the page lay-out (**?**).

This can easily be achieved using `endfloat.sty` package. Please refer the following documentation to use this package.

```
http://mirrors.ctan.org/macros/latex/contrib/
endfloat/endfloat.pdf
```

You should insert a caption for the figures below the figures and for the tables the caption should be above the tables.

Please remember that we will always also need highresolution versions of your artwork for printing, submitted as separate files in standard format (i.e. TIFF or EPS), not included in the text document. Before preparing your artwork, please take a look at our Web page: `http://www.elsevier.com/locate/authorartwork`.

3.2. *Lists*

For tabular summations that do not deserve to be presented as a table, lists are often used. Lists may be either numbered or bulleted. Below you see examples of both.

1. The first entry in this list
2. The second entry
   2..1 A subentry
3. The last entry

- A bulleted list item
- Another one

3.3. *Equations*

Conventionally, in mathematical equations, variables and anything that represents a value appear in italics. All equations should be numbered for easy referencing. The number should appear at the right margin.

$$S'_{\rm pg} = \frac{S_{\rm pg} - \min(S_{\rm pG})}{\max(S_{\rm pG} - \min(S_{\rm pG})} \tag{1}$$

In mathematical expressions in running text "/" should be used for division (not a horizontal line).

Acknowledgments

Acknowledgments should be inserted at the end of the paper, before the references, not as a footnote to the title. Use the unnumbered Acknowledgements Head style for the Acknowledgments heading.

References

Please ensure that every reference cited in the text is also present in the reference list (and vice versa).

Reference style

Text: All citations in the text should refer to:

1. Single author: the author's name (without initials, unless there is ambiguity) and the year of publication;
2. Two authors: both authors' names and the year of publication;
3. Three or more authors: first author's name followed by 'et al.' and the year of publication.

Citations may be made directly (or parenthetically). Groups of references should be listed first alphabetically, then chronologically.

References

Banerjee, A., Vogelstein, J.T., Dunson, D.B., 2013. Parallel inversion of huge covariance matrices. arXiv preprint URL: http://arxiv.org/abs/1312.1869, arXiv:1312.1869.

CHEN, S., BILLINGS, S.A., LUO, W., 1989. Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control 50, 1873--1896. URL: http://www.tandfonline.com/doi/abs/10.1080/00207178908953472, doi:10.1080/00207178908953472.

Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273--1302. URL: http://www.sciencedirect.com/science/article/pii/S1053811903002027, doi:10.1016/S1053-8119(03)00202-7.

Hsieh, C.J., Sustik, M.A., Dhillon, I.S., Ravikumar, P.K., Poldrack, R., 2013. BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables, in: Advances in Neural Information Processing Systems, pp. 3165--3173. URL: http://papers.nips.cc/paper/4923-big.

Johansen, S.r., 1988. Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control 12, 231--254. URL: http://www.sciencedirect.com/science/article/pii/0165188988900413, doi:10.1016/0165-1889(88)90041-3.

Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering 82, 35. URL: http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430402, doi:10.1115/1.3662552.

Kalman, R.E., 1963. Mathematical Description of Linear Dynamical Systems. Journal of the Society for Industrial and Applied Mathematics Series A Control 1, 152--192. URL: http://epubs.siam.org/doi/abs/10.1137/0301010, doi:10.1137/0301010.

Ljung, L., 1998. System Identification, in: Procházka, A., Uhlí, J., Rayner, P.W.J., Kingsbury, N.G. (Eds.), Signal Analysis and Prediction. Birkhäuser Boston, Boston, MA. Applied and Numerical Harmonic Analysis. URL: http://link.springer.com/10.1007/978-1-4612-1768-8, doi:10.1007/978-1-4612-1768-8.

Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257--286.

Xie, Y., Huang, J., Willett, R., 2013. Change-Point Detection for High-Dimensional Time Series With Missing Data. IEEE Journal of Selected Topics in Signal Processing 7, 12--27. URL: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6381435, doi:10.1109/JSTSP.2012.2234082.

Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., 2010. Spark: cluster computing with working sets, in: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, p. 10.

Supplementary Material

Supplementary material that may be helpful in the review process should be prepared and provided as a separate electronic file. That file can then be transformed into PDF format and submitted along with the manuscript and graphic files to the appropriate editorial office.