Compression-based and updating algorithms for constrained decompositions: explicitly and implicitly given tensors

A Data Management Plan created using DMPonline.be

Creators: Nico Vervliet, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Project Administrator: n.n. n.n.

Grant number / URL: 12ZM223N

ID: 198676

Start date: 01-10-2022

End date: 30-09-2025

Project abstract:

Almost all fields in science and engineering rely on data to analyze phenomena, predict events, discover hidden patterns, etc. While tools based on matrices have been very successful, only two modes are taken into account. In contrast, techniques for multiway arrays of numbers, or tensors, can handle three or more modes, often leading to more compact and more interpretable models.

We develop a proximal optimization algorithm for large-scale tensors which allows prior knowledge to be included, facilitating the discovery of meaningful components. When a tensor is only implicitly available as the solution of a linear system, its explicit construction can be undesirable or even impossible. Therefore, we rely on the blessing of dimensionality to solve the linear system and decompose its solution simultaneously. As these systems are typically large-scale, we develop algebraic and randomized compression algorithms enabling new applications in signal processing, data analysis, and machine learning.

When only limited processing time and storage capacity are available, incorporating new data in an existing tensor model has to be efficient in terms of memory and computational cost. We develop algorithms that decompose implicitly given tensors presented as a stream of data using a low memory footprint, and that update a decomposition when the underlying model changes or older data becomes obsolete, which are key problems in online and continuous monitoring applications.

Last modified: 21-04-2023

Compression-based and updating algorithms for constrained decompositions: explicitly and implicitly given tensors

Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

Random data will be generated using scripts to assess the performance of the various algorithms developed in this project. The data is therefore stored as a script with fixed random number generators.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

The STADIUS Data Management Register (DMR) is used to register the existence of a dataset and its metadata. Nico Vervliet, the owner of the data, is responsible during his employment at KU Leuven; afterward the responsibility moves to the promoter. A data management officer (Nico Vervliet) is assigned to guide this process.
During the research, random data is generated using scripts, which are stored on ESAT's internal GitLab server. When a work package is finished, the dataset moves to the STADIUS Dataset Server which is backed up regularly. An offline backup of this data is stored in the STADIUS Dataset Archive for 10 years after the project ends.

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

In the spirit of reproducible research, I will not deviate from this minimum preservation period. The STADIUS policy is that data is kept for 10 years after the project ends.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

No, the data will be randomly generated.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

In our research team, we aim to not only manage all data, but also to actively disseminate the data and the results. To this end, we released the Tensorlab+ project in December 2021. Tensorlab+ is a website where anyone can download all algorithms, scripts and data necessary to reproduce the results presented in journal papers, as well as tutorials and demos aimed at helping users to get started with the newly developed algorithms.

Compression-based and updating algorithms for constrained decompositions: explicitly and implicitly given tensors
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Compression-based and updating algorithms for constrained decompositions: explicitly and implicitly given tensors
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Compression-based and updating algorithms for constrained decompositions: explicitly and implicitly given tensors
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
mndot_sevd	main algorithms simultaneous eigenvalue decomposition for the (M,N,.) decomposition	new	digital	software	.m	<100MB	
mndot bsi	synthetic DS-CMDA data for blind system identification, including experiment files	new	digital	simulation data	.m	<100MB	
macaulay tensor	main algorithms for macaulay tensor method	new	digital	software	.m	<100MB	
macaulay test problems	example systems of polynomial equations	new	digital	simulation data	.m	<100MB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

n.a.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The algorithms developed during this project will be part of Tensorlab, a Matlab toolbox for tensor computations and complex optimization, developed and maintained in our research group. The toolbox is free to use for noncommercial, academic research as stipulated in the time-limited license. Commercial licenses can be obtained on request.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

In the research stage, algorithms and experiments are stored as matlab or python source files and are stored locally using version control software (GIT) and on ESAT's Gitlab server for backup and collaboration. Electronic notebooks with mathematical derivations, proofs and experiment details and results are stored as org-mode or latex files, which are plain text and can be viewed in any editor. These notebooks are stored locally and on ESAT's Gitlab server.

After publication of a paper, algorithms and experiments files will be standardized in terms of structure, documentation, readme, tutorials, etc. The data of this reproducible research version of the paper will be collected in a zip-file which will be made available via tensorlabplus.net, our reproducible research repository. These zip files will be made available as soon as possible. Relevant general purpose algorithms will be made available in Tensorlab as well but are released less often as several results are typically bundled in a bigger release.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

The results for each paper are made available via tensorlabplus.net, which contain a standardized html page per paper. Each page contains all bibliographic information as well as a zip file with all code and data needed to reproduce the results. The zip file is generated using an automated script which checks if all required elements (data, code, citation, readme, auxiliary files) are present and if all code runs in a clean environment. Templates have been generated for scripts, algorithms and readme files.

3. Data storage & back-up during the research project

Where will the data be stored?

During research, all data will be stored locally and on the internal Gitlab server from ESAT.

How will the data be backed up?

During the research project, the data will be backed up according to the standard ESAT procedure for user data on their servers.

When (part of) a project is finished, the data will be archived in the STADIUS Dataset Archive, which will be backed up online and offline.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

As the data is generated by scripts, only seeds for the random number generator and scripts (text files) need to be stored. For very time consuming experiments that cannot be reproduced quickly, the experimental results can be stored, but these typically require tens to hundreds of megabytes at most.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

As the data generated and used in this system is not sensitive, no specific security measures are required. The data stored in the ESAT servers has access regulated by an access control list (ACL) that grants: read-write access to the project owner read-only access to specific users The ACL is managed by the project owner. Client computers can access the data using: SMB2 (or higher) from specific IP ranges NFSv4 from specific (IT managed) systems.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Given the small size of the data, these cost are negligible as they fall within the standard user data allotment. The storage facilities of the research unit are available for the researchers for free.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All algorithms, code and data will be retained for 10 years after the project finishes per KU Leuven RDM policy.

Where will these data be archived (stored and curated for the long-term)?

During the research, random data is generated using scripts, which are stored on ESAT's internal GitLab server. When a work package is finished, the dataset moves to the STADIUS Dataset Server which is backed up regularly. An offline backup of this data is stored in the STADIUS Dataset Archive for 10 years after the project ends.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Given the small size of the generated data, the cost of storage is insignificant. The storage facilities of the research unit are available for the researchers for free.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)

The algorithms, code and experiments required to reproduce published results are made available via tensorlabplus.net. This repository is open access but requires registration and the user has to accept a license. This license is time-limited (one year) and allows using the results for academic, non-commercial research. For other types of use, a tailored license has to be requested via LRD.

If access is restricted, please specify who will be able to access the data and under what conditions.

A valid email address is needed to access the data as a download link is sent to the user. The user has to accept the time-limited license for academic, non-commercial use. Licenses for other types of use are discussed with LRD on a case-by-case basis.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

- https://www.tensorlabplus.net: all algorithms, code, experiment files and data.
- https://www.tensorlab.net: selected algorithms.

When will the data be made available?

- Data related to publications will be made available upon publication, or earlier if required for reviewing or collaboration purposes.
- A Tensorlab release will be planned if a sufficient number of new algorithms are available.

Which data usage licenses are you going to provide? If none, please explain why.

A time-limited license for academic, non-commercial research will be used. Licenses for other types of use are discussed with LRD on a case-by-case basis.
See https://tensorlabplus.net/license.html

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- No

What are the expected costs for data sharing? How will these costs be covered?

Given the small size of the datasets, the cost will be negligible: only a few 100 MB of storage on an accessible server (the ESAT web server) are required. The storage facilities of the research unit are available for the researchers for free.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Nico Vervliet

Who will manage data storage and backup during the research project?

Nico Vervliet

Who will manage data preservation and sharing?

Nico Vervliet

Who will update and implement this DMP?

Nico Vervliet