Article

Dataset Reduction Techniques to Speed Up SVD Analyses

Laurens Bogaardt [1] ⓘ **, Romulo Goncalves** [1] ⓘ **, Raul Zurita-Milla** [2,*] ⓘ **and Emma Izquierdo-Verdiguier** [2,3]

[1] Netherlands eScience Center; l.bogaardt@esciencecenter.nl, r.goncalves@esciencecenter.nl
[2] Faculty ITC, University of Twente; r.zurita-milla@utwente.nl
[3] Faculty IPL, Universitat de Valencia; emma.izquierdo@uv.es
[*] Correspondence: r.zurita-milla@utwente.nl

1 **Abstract:** An SVD is a mathematical procedure which has found multiple applications in the
2 geosciences. Performing SVD analyses on large datasets, however, can be computationally costly and
3 time consuming. Often, techniques exist to arrive at the same output, or at a close approximation,
4 which require far less effort. This article examines several such techniques in combination with
5 the inherent scale of the structure within the data. When the values of a dataset vary slowly, e.g.
6 in a spatial field of temperature over a country, the field contains large scale structure and there
7 is a high level of autocorrelation. Datasets do not need a high resolution to describe such fields.
8 Using simulated Gaussian Random Fields with various levels of autocorrelation, we examine rank
9 decomposition, coarsening and approximate SVD procedures. As the main result, this article provides
10 researchers with a decision tree indicating which technique to use when and predicting the resulting
11 level of accuracy based on the dataset's structure scale. Finally, these techniques and predictions,
12 based on simulated data, are verified using real-world geospatial datasets.

13 **Keywords:** Singular value decomposition, autocorrelation, rank deficiency, data reduction,
14 coarsening, approximate SVD, Gaussian Random Fields

1. Introduction

16 A *Singular Value Decomposition* (SVD) is a mathematical procedure which has found multiple
17 applications in the geosciences. When analysing a spatial field for a single time period, SVD's can help
18 identify structure within the data. It can also be used to filter out noise. In many real-world applications,
19 however, fields include multiple time periods spanning weeks, months or years. Researchers typically
20 compare two such datasets using *Maximum Covariance Analysis* (MCA) and *Canonical Correlation*
21 *Analysis* (CCA). These analyses rely on the SVD procedure to find patterns which occur frequently and
22 simultaneously [1,2].
23 Performing SVD's on large datasets can be computationally costly and time consuming. Often,
24 techniques exist to arrive at the same output, or at a close approximation, which require far less effort.
25 This article examines several procedures which exploit autocorrelation and rank decomposition to
26 analyse data in an efficient manner. We describe several SVD techniques, explain the type of problems
27 that could be tackled with them and make predictions about the error incurred in the approximations
28 based on the level of autocorrelation of the input data. To the best of our knowledge, this is the first
29 review in the geosciences which examines SVD implementations for large datasets from the point of
30 view of autocorrelation and data reduction. Even though these techniques are not novel, a review is
31 beneficial for domains less familiar with analysing large datasets [3–5].

To arrive at these predictions, *Gaussian Random Fields* (GRF's) are generated with various levels of autocorrelation and are subsequently reduced in size. The amount of error incurred in this reduction is determined by comparing the SVD of the reduced dataset to that of the original. Finally, the techniques and predictions, based on simulated data, are verified using real-world geospatial datasets. The reported results come from calculations performed in an accompanying *Jupyter Notebook* which can be found in the online supplementary material [6]. In order to develop intuition, some matrix algebra is briefly reviewed first, followed by discussion on the data characteristics of typical fields used in the geosciences.

1.1. Matrix Decomposition

Many datasets can be represented by a matrix; for instance, a group of n individuals who report scores on m questions or the temperatures at m locations measured over n time periods. These values can be arranged in a matrix with m rows and n columns. Like a vector, a matrix is a combination of basis vectors which indicate direction, each with a coefficient which indicates magnitude. As an extension of the vector, a matrix has two bases, the row- and the column basis, which can also be changed via a rotation. A clever basis to rotate into is one where the product of the first row- and column basis vectors explains as much of the variance in the dataset as possible. Subsequent pairs of basis vectors, known as *modes*, explain as much of the remaining variance as possible while being orthogonal to all previous modes. Such basis vectors are called *Principle Components* (PC's) or *Empirical Orthogonal Functions* (EOF's) and they are found via an SVD of the matrix.

If there exists a rotation for which some coefficients become zero, the matrix needs fewer basis vectors to describe it than are available. In a sense, it is underdetermined; its internal dimension is smaller than what would be expected from its m by n size. This is the concept of matrix *rank*; if the rows and columns both span a subspace of dimension r, a matrix has rank r. A matrix is said to have full rank if $r = \min(m, n)$, the maximum number of linearly independent basis vectors. It is rank deficient if $r < \min(m, n)$. A rank decomposition or factorization is the splitting of a matrix into a product where each factor has full rank. For example, an m by n matrix of rank r can be decomposed into an m by r matrix multiplied by an r by n one.

An SVD is a special type of rank decomposition which results in a set of orthonormal column basis vectors U, a list of coefficients s and a set of row basis vectors V. For rank deficient matrices, some of the coefficients, known as singular values, are zero. These zero singular values and their associated basis vectors can be truncated without affecting the data, which helps reduce storage and computational requirements. Whereas the entire, original dataset requires $m \times n$ units of storage and computing the product with a vector requires $m \times n$ flops, the rank decomposed version requires $m \times r + r \times n$ units of storage and an equal number of flops for the vector multiplication [7]. If r is small, this can be a substantial improvement. Thus, it is possible to use the rank decomposed version to reduce storage and have faster computation.

1.2. Data Characteristics

Real-world data are gathered by machines with finite precision, which means that the *information* contained in the data is limited by the noise level. Therefore, the mathematical rank r of a dataset is usually not relevant in practice [7]. Even though some singular values of a dataset are not zero, they may be small enough to be considered noise. If we take the inherent imprecise nature of real-world data into account, we can approximate a dataset by another matrix of rank l, with $l < r$, without losing much relevant information.

Following the Eckart-Young theorem, the best approximation is one described in the same bases as the original dataset, taking a subset of the l largest singular values and truncating the remainder [8]. Setting a threshold ϵ, the dataset is approximate rank deficient if some singular values fall below ϵ. Then, it has an ϵ-rank of l and the spectral norm of the difference with its approximation is at most ϵ [7]. The idea is to reduce the data to the point where the error due to reduction is around the noise level.

80   Beside irrelevant noise, real-world datasets also contain redundancy. In domains such as climate
81 science, datasets are typically spatio-temporal fields e.g. of global temperatures. In such fields, values
82 vary slowly and neighbouring points are not entirely independent of one another, neither in space nor
83 in time [1]. Then, there is a high level of autocorrelation and the field contains large scale structure.
84 Such redundancy in the data means the matrix is approximate rank deficient and that it can be reduced
85 in size without losing much relevant information. Using simulated spatio-temporal fields with different
86 levels of autocorrelation, we study the relation between autocorrelation and the accuracy of different
87 SVD implementations.

2. Materials and Methods

89   To compare our techniques and to establish a relation between performance and structure scale,
90 we need to be able to generate fields which resemble those often encountered in real-world applications.
91 Additionally, we require methods to control and measure the level of autocorrelation of a field.

2.1. Artificial Spatio-Temporal Fields

93   As simulated spatio-temporal fields, real-valued *Gaussian Random Fields* (GRF's) are particularly
94 useful because their structure scale can be captured in a single parameter. For such rotational invariant
95 fields, the spectrum follows the power law described by $P(\vec{k}) \sim |\vec{k}|^{-\alpha}$ where $\vec{k}$ is the wavevector and α
96 the parameter which controls the level of autocorrelation. Figure 1 shows fields with various α's.

(a) $\alpha = 1$ **(b)** $\alpha = 3$ **(c)** $\alpha = 5$
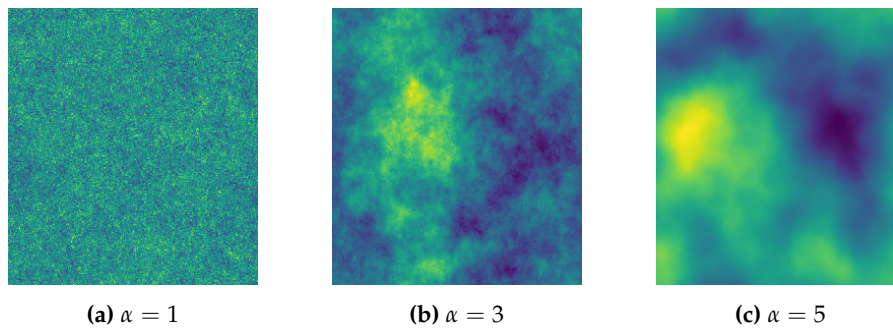
**Figure 1.** Gaussian Random Fields for various α's

97   Just as there is spatial autocorrelation, there is temporal autocorrelation, when the values of the
98 field over the entire time period do not change drastically. In principle, there can be different levels of
99 autocorrelation over time and over space. However, for simplicity, in this article we will use the same
100 level of autocorrelation in all directions, determined by parameter α.

2.2. Autocorrelation

102   In the geosciences, there are additional measures of spatial autocorrelation [1,2]. One frequently
103 used is Moran's I [9–11]. Figure 2 shows the relationship between Moran's I, using a uniform kernel
104 with a bandwidth equal to 10, and the α of our simulated GRF's.
105   One can also devise an autocorrelation measure from the singular values of a dataset. Each
106 singular value indicates the amount of variance explained by its associated mode. For fields with
107 autocorrelation, the sorted list of singular values decays quickly. A power law can be fitted to this list,
108 with an exponent which we call β.
109   A high α implies a high Moran's I and a high β, which, in turn, implies some singular values are
110 close to zero. Therefore, spatial fields with high levels of autocorrelation are described by matrices
111 which are approximate rank deficient. This allows for data reduction without losing much information.
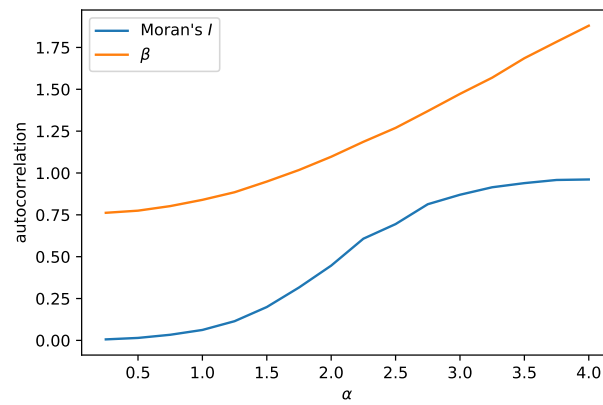
Figure 2. Measures of autocorrelation as a function of α

3. Results

This section lists several SVD related implementations to analyse large datasets efficiently by exploiting autocorrelation and rank deficiency. Additional methods exist, though the ones here cover three main benefits: the product SVD via QR decomposition provides an exact result, dimensionality reduction provides the best possible approximation for any level of reduction, while coarsening is an easy-to-implement method. Which technique to use when is described in the decision tree of figure 3.

There are several reasons for performing an SVD analysis. When applied to a single spatial field, it may be to find the PC's or EOF's, which describe areas that behave similarly. In many real-world applications, however, the analysis of a field does not only involve a single time period but includes data over multiple weeks, months or years. Researchers typically compare two such datasets using MCA and CCA to find patterns which occur frequently and simultaneously [1,2]. Such a pattern, or mode, is a combination of a row- and a column basis vector. One technique to determine these modes is to perform an SVD on the product of the standardised datasets. In some domains, the term SVD is used synonymously with MCA. In an MCA, modes are found where the row- and the column vector covary maximally, whereas in a CCA, they correlate maximally [12].

To help researchers identify situations where different SVD approaches are beneficial, we have constructed a decision tree in figure 3. The first question to be answered is whether the SVD will be applied to a single matrix or to the product of two matrices. For single fields, the data may be small enough to fit in the memory of a computer. Then, a regular SVD is the best option. If the dataset is too large, two alternatives exists which provide an approximate answer: coarsening and randomised dimensionality reduction. These will be discussed in section 3.1 and section 3.2.
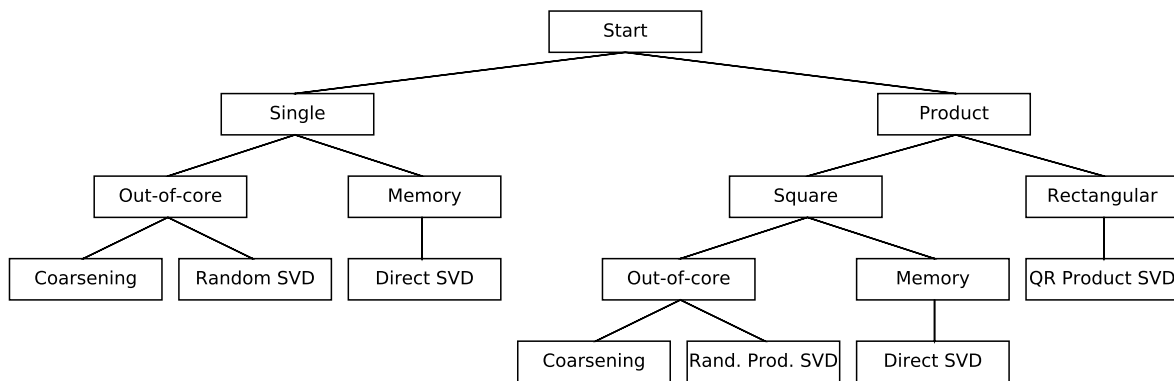
Figure 3. Decision tree describing the possible SVD techniques

133 When the SVD is performed on the product of two matrices, the best course of action depends on
134 whether the matrices are square or rectangular. The rank of a matrix is at most the size of the smallest
135 side, which, for rectangular matrices, can be small. How to exploit this fact is described in section 3.7.
136 Square matrices small enough to fit in memory can be analysed directly. Variations of coarsening and
137 dimensionality reduction can assist analyses of larger datasets, discussed in section 3.4 and section 3.5.

138 *3.1. Approximate SVD of a Single Matrix via Coarsening*

139 When a spatial field has large scale structure, the values of neighbouring cells do not change
140 drastically. Perhaps these cells can be aggregated together to produce a smaller dataset which still
141 faithfully describes the original field. Here, we coarsen two dimensional GRF's.
142 Figure 4 shows the error in a coarsening process for matrices of various α's and for different
143 coarsening window sizes. The error is determined as the norm of the difference between the original
144 matrix and the coarsened version, divided by the norm of the original [6]. Other measures of similarity,
145 such as the correlation between the two datasets, are likely to show the same effects. Note that
146 coarsening a two dimensional field at level 5 reduces the dataset by 25 times. For fields with high
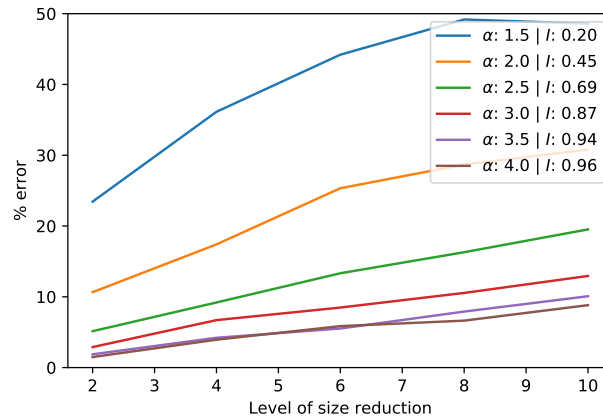147 autocorrelation, e.g. $\alpha = 3$, the coarsened version differs by less than 10% from the original.

Figure 4. Error in the SVD of a coarsened spatial field for various α's

148 *3.2. Approximate SVD of a Single Matrix via Dimensionality Reduction*

149 An alternative method of reducing the size of a matrix is via randomised dimensionality reduction.
150 During dimensionality reduction, the number of basis-vectors of a matrix is truncated to the l most
151 important ones, similar to finding an ϵ-rank approximation. Remember from section 1.1 that matrices
152 with low rank require less storage space and fewer flops for vector multiplications. A random
153 algorithm exists which performs an approximate rank decomposition of a large matrix efficiently. The
154 mathematics behind this algorithm is reviewed extensively elsewhere, here we examine its performance
155 on datasets with autocorrelation [13,14]. The main idea is to sample from the original values, taking
156 as many values as needed to achieve the required accuracy. One of its benefits is that this algorithm
157 requires only a small number of passes over the data, which, for large matrices stored out-of-core,
158 reduces reading time. Additionally, the incurred error in the approximation can be made arbitrarily
159 small by adjusting l and ϵ, giving the researcher full control over the balance between computation
160 cost and accuracy.
161 Figure 5 depicts the calculation in this process, which first reduces the input matrix to a smaller
162 square matrix of l by l. It also provides two projection matrices which rotate the rows and columns of
163 this smaller matrix back as close as possible to the bases of the original input. Subsequently, the SVD is
164 applied to the small l by l matrix, which results in a fast and efficient approximation of the matrix's
165 decomposition with an error of the order of the size of the largest truncated singular value [7,13].

$$A \approx H \boxed{L} \boxed{W^T} = H \boxed{\breve{U}} \boxed{S} \boxed{\breve{V}^T} \boxed{W^T} = U \boxed{S} \boxed{V^T}$$

Figure 5. Visualising the calculation of an approximate SVD via dimensionality reduction

166   The calculations in the *Jupyter Notebook*, which is provided in the suplementary material, show
167 that the errors induced by the randomised SVD procedure are very small, even for high levels of
168 reduction [6]. This is especially true for fields with autocorrelation. Looking at figure 2, this is not
169 surprising; high autocorrelation means a high β which implies the singular values decay quickly and
170 that the last modes contribute little information to the dataset.

171   This randomised technique achieves much lower errors than coarsening. The coarsening
172 procedure has several advantages though. For one, it is intuitive and the results are easy to interpret.
173 It is also trivial to implement. Additionally, different coarsening levels can be applied to different
174 directions. This is especially advantages when directions have different levels of autocorrelation or are
175 recorded at different resolutions. Finally, the predictions of figure 4 can help researchers determine at
176 what resolution to gather their data in the first place. In domains were satellite data is used, datasets
177 are often not very detailed because the imaging resolution is low. Unlike local analyses of developed
178 countries, where high resolution data is becoming more accessible, for continental or global analyses,
179 coarse spatial resolution data may simply be the only option.

3.3. Case Study of an SVD of a Single Matrix

181   Coarsening and dimensionality reduction are particularly useful for spatial fields with high
182 levels of autocorrelation. As an example of this, we examine humidity and cloud cover data from the
183 ERA5 datasets for a single time period measured on 1 April 2012. ERA5 is an atmospheric reanalysis
184 of the global climate using high spatial resolution forecasts, produced by combining models with
185 observations [15]. It contains estimates of atmospheric parameters such as air temperature, pressure
186 and wind at different altitudes.

187   Clearly, the ERA5 humidity and cloud cover fields are not GRF's. Nonetheless, we can get an idea
188 of the accuracy of an SVD after data reduction if we estimate the levels of autocorrelation. This can
189 verify whether simulated GRF's are reasonable representations of real-world datasets. The humidity
190 field has a Moran's $I \approx 0.98$, while the cloud cover data shows less structure with a Moran's $I \approx 0.82$.
191 The estimations for α were unreliable as the power law did not fit properly, though figure 2 can help
192 us translate the measures and suggests the fields are equivalent to GRF's with an $\alpha \sim 3.5$ and $\alpha \sim 2.5$,
193 respectively.

194   The coarsening predictions of figure 4 indicate the first field is expected to incur errors around
195 a few percent for size reductions between 2 and 8, while for the second field we should see errors
196 between 5% and 15%. The calculations in the *Jupyter Notebook* in the supplementary material show
197 that this prediction is fairly accurate, perhaps slightly pessimistic. The humidity field, which has the
198 largest autocorrelation, only incurred an error of 1% when halved in size. For the cloud cover data, this
199 values was closer to 4%. Likewise, we can apply the randomised dimensionality reduction technique
200 to the fields. As expected, this resulted in very small errors, below 1%. If a researcher cares most about
201 accuracy, dimensionality reduction is the best option.

3.4. Approximate Product SVD of Square Matrices via Coarsening

203   In real-world applications, researchers often wants to find the relation between two fields.
204 Analyses such as the MCA and CCA, discussed in section 1, rely on performing an SVD of the product
205 matrix of two spatio-temporal fields. If the input datasets have the various spatial gridpoints as rows
206 and the sample of recorded values over time as columns, multiplying them gives their cross-covariance

207 matter. An SVD of this cross-covariance matrix exposes the patterns in the data which occur frequently
208 and simultaneously. Such a matrix product can, however, become rather large and difficult to analyse.
209   In section 3.1, we saw coarsening can reduce the size of a single spatial field. We can also
210 coarsen two spatio-temporal fields before analysing their cross-covariance matrix. Figure 6 shows the
211 percentage error for various simulated spatio-temporal fields. Note that only the spatial directions are
212 coarsened in our calculation. This is because the time direction gets consumed in the matrix product
213 of the MCA or CCA and coarsening it will not decrease the size of the cross-covariance matrix nor
214 speed up the SVD. While coarsening two spatial directions means each field is reduced by the square
215 of the coarsening level, the cross-covariance matrix is reduced by this level to the power 4. As a
216 result, the typical error in this product is larger than for the single field, though the speed up is also
217 substantial. Clearly, the level of autocorrelation plays an important part, with larger α's leading to less
218 error. Preliminary work shows the amount of error also depends on the similarity between the two
219 spatio-temporal fields, though we leave deeper investigation of this aspect for further research.
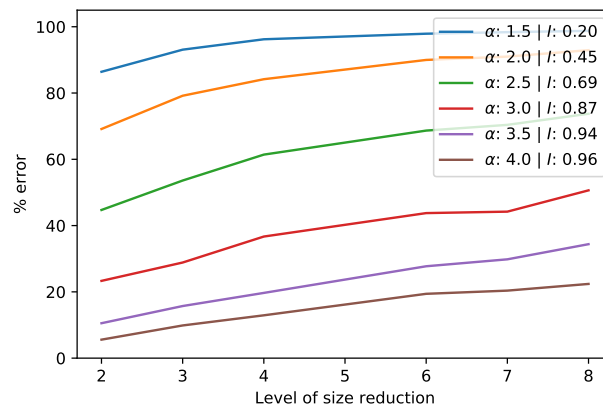
Figure 6. Error in the SVD of the product of two coarsened fields for various α's

220 *3.5. Approximate Product SVD of Square Matrices via Dimensionality Reduction*

221   The randomised dimensionality reduction process may also speed up the SVD analysis of two
222 spatio-temporal fields. The reduction can be applied to each of the fields before they are multiplied
223 into the cross-covariance matrix. Similar to the QR decomposition of section 3.7, it has the advantage
224 that the SVD is performed on a small l by l matrix. This calculation is visualised in figure 7. In the first
225 step, the input datasets are decomposed using the algorithm discussed in section 3.2 and reviewed
226 extensively elsewhere [13,14]. Again, the main idea behind the algorithm is to sample from the original
227 values, taking as many values as needed to achieve the required accuracy. Subsequently, the inner
228 matrices are multiplied and the SVD is performed on the resulting, small matrix.

Figure 7. Visualising the calculation of an approximate SVD for the product of two fields using dimensionality reduction

229   After generating various GRF's, the reduced cross-covariance matrix is compared with the
230 original. Figure 8 shows that the results are not satisfactory for fields with a small α, but high levels of
231 autocorrelation allow for substantial savings in computation time without incurring much error. Note
232 that, when performing an MCA or CCA on a spatio-temporal field, the spatial directions are flattened
233 and some of the spatial autocorrelation is lost. This partially explains why the error is substantial for
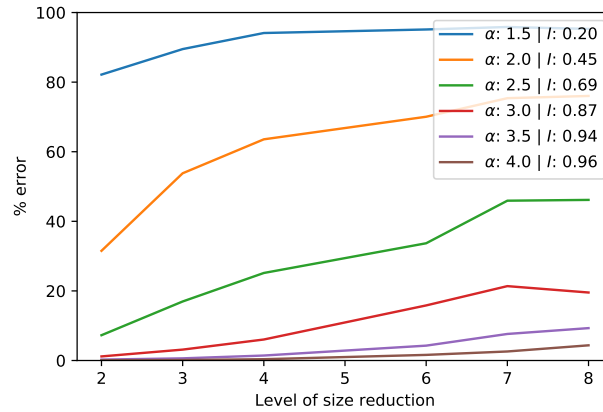234 low α.

Figure 8. Error in the SVD of the product of two reduced fields for various α's

The reduction of the number of dimensions of each input dataset before an MCA or CCA is actually advised by some researchers, as a method to filter out noise [16]. Especially when the number of temporal samples is small, outliers and random fluctuations could affect the result [12]. This is because any statistical analysis will choose its regression-coefficients so as to optimize the fit. It may occur that two noise-vectors in the two fields coincidentally covary and show up as dominant modes. Prefiltering can alleviate this risk.

3.6. Case Study of a Product SVD of Square Matrices

The JRA55 data is an atmosphere reanalysis product which includes quantities such as humidity, pressure and temperature [17]. Recently, these quantities were used to determine the total meridional energy transport and latent heat, measures important to understand the global climate [18]. As an example of our reduction techniques for matrix products, we are using a Mercator projection of the energy transport and latent heat, recorded monthly from 1979 to 2015. Although the spatially flattened matrices are not completely square, their high resolution in the time direction make them substantially less rectangular than the phenology data from section 3.8. Therefore, this serves as a good use case for the coarsening and dimensionality reduction techniques for a square product SVD.

The energy field has a Moran's $I \approx 0.93$, while the latent heat field has a Moran's $I \approx 0.86$. The estimations for α were unreliable, though figure 2 can help us translate the measures and suggests the fields are equivalent to GRF's with an $\alpha \sim 3.0$ and $\alpha \sim 2.5$, respectively. The analyses of section 3.4 and of section 3.5 make the prediction that, for such α levels, our data reduction techniques result in errors which are quite high. In contrast, we found that coarsening the fields before applying an SVD on their product merely resulted in an error of 8% when the data was halved and an error around 16% when the data was reduced by a factor of 4. For the dimensionality reduction technique, as well, the predictions overstated the observed error, which were around a few percent, even for high levels of size reduction.

3.7. Exact Product SVD of Rectangular Matrices via QR Decomposition

In an MCA and CCA, an SVD is performed on the product of two spatio-temporal fields to find patterns which occur frequently and simultaneously [1,2]. For highly rectangular datasets, when there are many spatial gridpoints but few temporal samples, the resulting cross-covariance matrix is inefficiently large and obviously rank deficient. Performing a rank decomposition on each dataset before multiplying them allows the SVD to be calculated in an efficient manner [5,19]. Figure 9 depicts the calculation in this technique. Using the QR decomposition, the input datasets are first transformed into two small, square matrices together with rectangular orthonormal basis vectors. The SVD is then performed on the product of the square matrices, giving a mathematically identical result to the full SVD while never forming an unnecessarily large, intermediate matrix.

Figure 9. Visualising the calculation of the exact SVD of a product via QR decomposition

3.8. Case Study of a Product SVD of Rectangular Matrices

Let's apply the QR decomposition technique to phenological datasets. Phenology is the science that studies recurring biological events such as leafing and blooming as well as their causes and variations in space and time. Spatio-temporal fields of remotely sensed images can be used to derive various phenological metrics. One important family of such metrics are the *Extended Spring Indices* (SI-x), which are based on a suite of models that transform daily temperatures into consistent phenological metrics [20]. In particular, we take versions of the Leaf and the Bloom index which were recently generated for the US by adapting the SI-x models to a cloud computing environment [21]. Both datasets span from 1989 to 2014 and have a 1km^2 spatial resolution, meaning there are far fewer time periods than spatial gridpoints, giving highly rectangular matrices. In fact, each dataset contains 30 million rows and 26 columns and is about 3.1 GB large. Their cross-covariance would be 30 million rows by 30 million columns and about 3.6 PB in size. Luckily, this product matrix need not be created to perform the SVD.

For the SVD via QR decomposition, the level of autocorrelation is irrelevant because this technique provides a mathematically exact result. Indeed, the *Jupyter Notebook* in the supplementary material, as well as work being prepared for publication, shows that this technique provides the full SVD of the cross-covariance matrix for these datasets in a matter of seconds, without ever exceeding the *RAM*-memory [6,22].

4. Discussion

4.1. Further Work and Caveats

Much of the analysis here relies on knowledge of the level of autocorrelation, which may be difficult to determine for large datasets. The *Jupyter Notebook* accompanying this article includes an algorithm which estimates Moran's *I* based on a sample of gridpoints. This speeds up the calculation substantially compared with the full calculation. Further work could be placed into making this algorithm more professional and more user friendly. Additional areas of research include relaxing the assumption that the autocorrelation in the time direction is similar to that in the spatial directions. In fact, it is more realistic to allow for different levels of autocorrelation in all directions and to have a version of Moran's *I* which can estimate these values.

Furthermore, a warning about autocorrelation and standardisation. In MCA's, the timeseries of each spatial gridpoint is centred about its mean and in CCA's, each gridpoint is standardised. These operations destroy much of the spatial autocorrelation, as it can affect neighbouring cells differently. When researcher choose the coarsening technique, this should occur before any additional data processing steps.

Unlike the coarsening procedure, the dimensionality reduction is not applied on each spatial field for each time period, but rather on the entire spatially flattened timeseries. Therefore, the level of spatial autocorrelation may not be as important as the level of temporal autocorrelation. Further work can examine how to apply the reduction to the spatial part of the spatio-temporal fields, before it is flattened. Alternative solutions, which retain the spatial structure, may include 3D tensor operations such as *Higher-Order Singular Value Decomposition* (HOSVD) [23].

4.2. Summary

In summary, randomised dimensionality reduction works best for datasets which are too large for internal memory. It requires only a small number of passes over the data, which decreases storage reading time. It also allows the researcher to balance computation cost with accuracy, by tuning the algorithm's parameters. Performing analyses at a coarse level can be beneficial when data collection is difficult and provides an intuitive and easy-to-implement alternative. These techniques require at least some autocorrelation in the fields, which results in approximate rank deficient datasets. In general, rank decompositions can speed up calculations by splitting datasets into smaller matrices of full rank. For the product of rectangular matrices, this can even give a mathematically exact result.

Abbreviations

The following abbreviations are used in this manuscript:

SVD	Singular value decomposition
GRF	Gaussian random field
PC	Principle component
EOF	Empirical orthogonal function
SOS	Start of season
SI-x	Extended spring indices
ERA5	European fifth generation reanalysis
JRA55	Japanese 55-year reanalysis
HOSVD	Higher-order singular value decomposition

1. Eshel, G. *Spatiotemporal data analysis*; Princeton University Press, 2011.
2. von Storch, H.; Zwiers, F.W. *Statistical analysis in climate research*; Cambridge University Press, 1999.
3. Golub, G.H.; Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* **1970**, *14*, 403–420. doi:10.1007/BF02163027.
4. Björck, Å.; Golub, G.H. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **1973**, *27*, 579–594.
5. Chan, T.F. An improved algorithm for computing the svd. *ACM Trans. Math. Softw.* **1982**, pp. 72–83. doi:10.1145/355984.355990.
6. Bogaardt, L. Dataset reduction techniques to speed up svd analyses. https://github.com/phenology/, 2018.
7. Martinsson, P.G. Randomized methods for matrix computations and analysis of high dimensional data. *ArXiv* **2016**.
8. Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, pp. 211–218. doi:10.1007/BF02288367.
9. Moran, P.A.P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23.
10. Hubert, L.J.; Golledge, R.G.; Costanzo, C.M. Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis* **1981**, *13*, 224–233. doi:10.1111/j.1538-4632.1981.tb00731.x.
11. Rey, S. PySAL. http://pysal.readthedocs.io, 2009–2013.

12. Bretherton, C.S.; Smith, C.; Wallace, J.M. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate* **1992**, *5*, 541–560. doi:10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2.

13. Halko, N.; Martinsson, P.G.; Tropp, J.A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* **2011**, *53*, 217–288. doi:10.1137/090771806.

14. Li, H.; Kluger, Y.; Tygert, M. Randomized algorithms for distributed computation of principal component analysis and singular value decomposition. *CoRR* **2016**, *abs/1612.08709*.

15. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; Bechtold, P.; Beljaars, A.C.M.; van de Berg, L.; Bidlot, J.; Bormann, N.; Delsol, C.; Dragani, R.; Fuentes, M.; Geer, A.J.; Haimberger, L.; Healy, S.B.; Hersbach, H.; Holm, E.V.; Isaksen, L.; Kallberg, P.; Köhler, M.; Matricardi, M.; McNally, A.P.; Monge-Sanz, B.M.; Morcrette, J.J.; Park, B.K.; Peubey, C.; de Rosnay, P.; Tavolato, C.; Thepaut, J.N.; Vitart, F. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **2011**, *137*, 553–597. doi:10.1002/qj.828.

16. Barnett, T.P.; Preisendorfer, R. Origins and levels of monthly and seasonal forecast skill for us surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review* **1987**, *115*, 1825–1850. doi:10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2.

17. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H.; Miyaoka, K.; Takahashi, K. The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan* **2015**, *93*, 5–48. doi:10.2151/jmsj.2015-001.

18. Liu, Y.; Attema, J.; Moat, B.; Hazeleger, W. Synthesis and evaluation of historical meridional heat transport from midlatitudes towards the arctic. *Climate Dynamics* **2018**, *Submitted*.

19. Tygert, M. Suggested during personal communication, 2017.

20. Schwartz, M.D.; Ault, T.R.; Betancourt, J.L. Spring onset variations and trends in the continental united states: past and regional assessment using temperature-based indices. *International Journal of Climatology* **2013**, pp. 2917–2922. doi:10.1002/joc.3625.

21. Izquierdo-Verdiguier, E.; Zurita-Milla, R.; Ault, T.R.; Schwartz, M.D. Using cloud computing to study trends and patterns in the extended spring indices. *Third International Conference on Phenology* **2015**, p. 51.

22. Zurita-Milla, R.; Bogaardt, L.; Izquierdo-Verdiguier, E.; Gonçalves, R. Analyzing the cross-correlation between the extended spring indices and the AVHRR start of season phenometric. *EGU General Assembly: Geophysical Research Abstracts* **2018**.

23. Tucker, L.R. The extension of factor analysis to three-dimensional matrices. In *Contributions to mathematical psychology*; Gulliksen, H.; Frederiksen, N., Eds.; Holt, Rinehart and Winston, 1964; pp. 110–127.