Information Loss During Size Reduction Depending On Structure Scale

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

[1]NLeSC Amsterdam, The Netherlands {*{l.bogaardt, r.goncalves}@esciencecenter.nl*}

[2]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, the Netherlands

{{*{r.zurita-milla,e.izquierdoverdiguier}@utwente.nl*}}

[3]Image Processing Laboratory (IPL), Universitat de Valencia, Spain

ABSTRACT

The analysis of large datasets can be time consuming and costly. Often, techniques exist to arrive at the same output, or at a close approximation, which require far less effort. This article looks at several such techniques and at the inherent scale of the structure within the data. When the values of a dataset vary slowly, e.g. in a spatial field of temperature over a country, there is a high level of autocorrelation and the structure of the field has a large scale. Datasets need not have a high resolution to describe such fields faithfully. Using generated *Gaussian Random Fields* with various levels of spatial autocorrelation, we examine several exact and approximate analysis techniques. Our aim is to outline when certain techniques can be useful and to find a relation between performance and the scale of the structure described by the input datasets.

1. INTRODUCTION

This article looks at several techniques to analyse large datasets and at the inherent scale of the structure within the data. When a researcher has an idea about the scale of their data and has a target level of accuracy, this article can suggest which technique may be relevant for their analysis. The techniques discussed here are by no means novel [9, 2, 6]. However, there are domains which are less familiar analysing large datasets, so a review may be beneficial.

1.1 Matrix Size and Rank Decomposition

Many datasets can be represented by a matrix of values. Take, for instance, a group of n idividuals who report scores on m different questions. Or take the temperatures at m locations, measured over n time periods. These values can be placed into 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 left- and the right-, or the row- and the column basis. Similarly, these bases can be changed via a rotation. Then, the coefficients will also change, leaving the resulting matrix untouched. A clever basis to rotate into is one where the bases are orthonormal and each subsequent set of left- and right basis vectors explains as much of the remaining variance in the dataset as possible. Such basis vectors are called *Principle Components* or *Empirical Orthogonal Functions* and they may be found via a *Singular Value Decomposition* (*SVD*) of the matrix.

If one can find a rotation in which one of the coefficients becomes zero, the matrix seems to be able to be described by fewer parameters than are available. In a sense, it is underdetermined. Its internal dimension is smaller than what could have been guessed from its m by n size. This is the concept of matrix 'rank'. An m by n matrix has rank r if the rows and the columns both span a subspace of dimension r. If $r = \min(m, n)$, such a matrix is said to have full rank, the maximum number of linearly independent basis vectors. If $r < \min(m, n)$, it is rank deficient.

A rank decomposition or factorization is the splitting of a matrix into a product where each factor has full rank. For an m by n matrix of rank r, with $r \leq n \leq m$, we can decompose it into an m by r matrix multiplied by an r by n one. Furthermore, we can choose the first factor of this product to be an orthonormal matrix which induces a rotation, i.e. a change-of-basis. The second factor captures the 'action' of the matrix, written in the new bases. It is this second matrix, which is often smaller than the original, which is most relevant for further analyses. An *SVD* is a special type of rank decomposition. It results in a set of orthonormal left basis vectors U, a list of coefficients s and a set of right basis vectors V. For rank deficient matrices, some of the coefficients, called singular values, will be zero.

As noted by Martinsson, the condition that a dataset has precisely rank r is not realistic in practice because the values originate from devices with finite precision [16]. Even though some singular values of a dataset are not zero, they may be close enough to zero to be considered *noise*. If we take the inherent imprecise nature of real-world datasets into account, we can approximate a dataset by another matrix of rank l, with $l < r$. Following the Eckart-Young-Mirsky theorem, the best possible 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 [7]. Taking a threshold ϵ, the dataset is said to be approximate rank deficient if some singular values fall below ϵ. Then, it has an ϵ-rank of l and the norm of the difference with its l-rank approximation is at most ϵ [16].

So, we can identify three types of matrix 'sizes'. The first is the size of the full matrix, $m \times n$. Storing such a matrix requires $m \times n$ units of storage and computing the product with a vector requires $m \times n$ flops. The second type is the rank decomposed version of the matrix. Storing such a matrix requires $m \times r + r \times n$ units of storage and an equal number of flops for the vector multiplication [16]. If r is small, this can be a substantial improvement. The final definition of 'size' approximates the original dataset with a matrix of rank l, resulting in even smaller storage and faster computations, while losing as little information as possible.

1.2 Spatial Fields

In domains such a climate science and phenology, datasets are typically spatial fields, e.g. of temperature. In these fields, values vary slowly and neighbouring points are not entirely independent of one another, neither in space nor in time [8]. Then, there is a high level of autocorrelation and the field has large scalestructure. This redundancy means the dataset is rank deficient. In this article, we will examine various fields and exploit rank factorization to analyse the data in an efficient manner. The reported results come from calculations performed in an accompanying *Jupyter Notebook* [3]. Note that the code was not optimised for speed, but merely serves to illustrate the procedures discussed here.

In order to compare our techniques and to find a relation between performance and structure scale, we need to be able to generate fields which resemble those often encountered in real-world applications. In particular, we will concern ourselves with fields which combine some level of autocorrelation with some randomness. Real-valued *Gaussian Random Fields* are particularly useful, as their structure scale can be captured in a single parameter. The spectrum of such fields follows the power law described by $P(k) = c_0\, k^{-\alpha}$ where k is the wavenumber and α the parameter which controls the level of autocorrelation. For 2D spatial fields, rotational invariance is assumed, such that k can be substituted by $|\vec{k}|$.

In spatial data analysis, other measures of autocorrelation are often used [8, 22]. These include Moran's I and the Γ index [17, 11, 19]. Another measure comes from the singular values. These are related to the amount of variance in the original dataset explained by their associated mode. For fields with autocorrelation, the singular values decay quickly. One can try to fit a power law to them and estimate the exponent, which we'll call β. All the measures give an indication of the level of autocorrelation in the field and the scale of the structure represented in the data. Figure 1 plots them as a function of α for various generated Gaussian Random Fields.

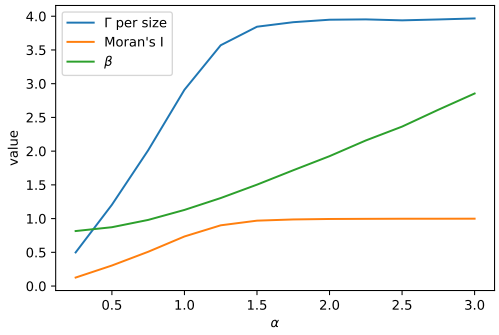

Figure 1: Various measures of autocorrelation as a function of α

1.3 Spatial-Temporal Fields

In many real-world applications, the analysis of a field does not only involve a single time snapshot but includes data over multiple weeks, months or years. Then, we are interested in finding patterns which occur frequently. The *maximum covariance analysis* and *canonical correlation analysis* examine the cross-covariance matrix of two datasets and find patterns which occur frequently and simultaneously [8, 22]. Such a pattern, or mode, is a combination of a left- and a right basis vector. One technique to find 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 left- and the right-field covary maximally, wereas in an CCA, they correlate maximally [4].

Just as there is spatial autocorrelation, there is temporal autocorrelation when the values of the field over the entire time period do not change drastically. In principle, there can be a different level of autocorrelation over time and over space. However, for simplicity, in this article we will use the same α to determine the level of autocorrelation in all dimensions.

2. TECHNIQUES

This section will discuss four techniques to analyse large datasets efficiently using their singular value decomposition and by exploiting autocorrelation and rank deficiency.

2.1 Exact Norm of Difference via SVD

One often wants to find the norm of the difference between two fields. This can be done by subtracting one matrix from the other and summing the square of the elements. However, for large matrices, this may be inefficient, especially when they are rank deficient and when their SVDs are already known.

Let $||\cdot||$ indicate the Frobenius norm, $\langle\cdot\rangle$ the Frobenius inner product and the $\circ$ operator the Hadamard product, then the norm of the difference between matrices A and B is given by equation 1.

$$\begin{aligned} ||A-B||^2 &= ||A||^2 + ||B||^2 - 2\langle A, B\rangle \\ &= s_A^T s_A + s_B^T s_B - 2 s_A^T \left(U_A^T U_B \circ V_A^T V_B\right) s_B \end{aligned} \tag{1}$$

Figure 2 shows that this procedure can determine the norm in an efficient manner, provided the number of singular values is small. The result is mathematically identical to the full calculation, which means that any error will be of the order of machine-precision.

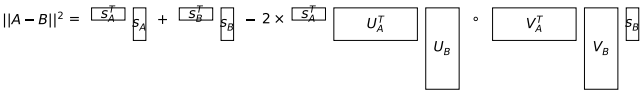

Figure 2: Exact norm of difference via SVD

2.2 Exact SVD via QR Decomposition

In real-world applications, one often wants to find the relation between two fields. Analyses such as the MCA and CCA discussed in section 1.3 rely on performing an SVD of the cross-covariance matrix of the two fields.Take wo input datasets with the various spatial gridpoints as rows and the sample of recorded values over time as columns. Centering and multiplying these gives the cross-covariance matrix. However, for highly rectangular matrices, when

there are many spatial gridpoint but few temporal samples, the resulting cross-covaraince matrix is inefficiently large and obviously rank deficient. Performing a rank decomposition, such as the *QR Decomposition*, allows one to do the SVD in an efficient manner [6, 21]. The result is mathematically identical to the full SVD, which means that the difference will be at machine-precision.

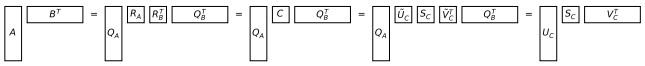

Figure 3: Exact SVD of a matrix product via QR decomposition

2.3 Approximate SVD via Spatial Coarsening

Although the previous procedure works well for two rectangular matrices, sometimes the input data is large and square. Performing an SVD on such large datasets will be time consuming and possibly inefficient given the desired level of accuracy. When a spatial field has large scale structure, the values of neighbouring cells do not change drastically. Perhaps these cells can be aggragated together to produce a smaller dataset which still faithfully describes the original field. In this section, we coarsen various Gaussian Random Fields by averaging patches of neighbouring gridpoints. We then compare the result with the full calculation.

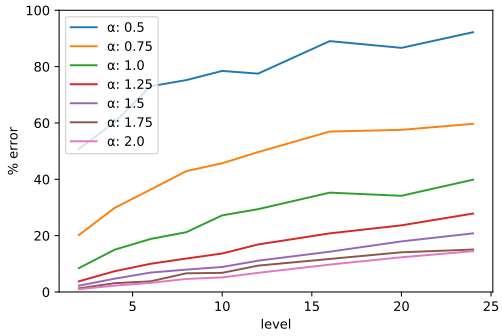

Figure 4: Error after coarsening a spatial field for various α's

Figure 4 shows the percentage error in a coarsening process for matrices of various sizes and α's, and at various levels of coarsening. The error is determined as the norm of the difference between the original matrix and the coarsened version, divided by the norm of the original.

We can also coarsen two different fields before analysing their cross-covariance matrix. Figure 5 shows the percentage error for various generated matrices. Due to the multiplication step in this analysis, the typical error as a result of coarsening is larger than before. As expected, the level of autocorrelation plays an important part, with more negative α's leading to a smaller error. The amount of error during the coarsening process will likely also depend on the similarity between the two datasets. This is one aspect which we do not cover here and leave for further research.

The coarsening process can speed up the calculation of the SVD, but there are additional benefits. When a target level of accuracy is determined, and there is an a priori estimate of the level of autocorrelation of the fields, the data collection process can be optimised. Knowing in advance at what resolution to gather data can help save time. Furthermore, in domains were satellite data is used, datasets

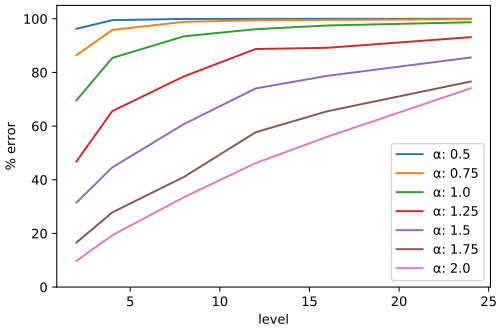

Figure 5: Error after coarsening the product of two fields

are often not very detailed because the imaging resolution is low. Unlike local analyses of developed countries, where high resolution data is becoming more accessible, for continental or global analyses, coarse spatial resolution data may simply be the only option.

2.4 Approx. SVD via Dimension Reduction

The spatial coarsening process is intuitive and easy to implement. It is not, however, the most efficient way to reduce the size of a dataset. Dimension reduction refers to discarding modes which contribute little to the variance in a dataset. As mentoned in section 1.1, an SVD is precisely the procedure used to find modes which explain as much variance as possible. Discarding the smallest singular values, therefore, gives the best lower rank approximation [7, 16]. Performing an SVD on a large dataset, however, is computationally costly. The *Randomised Dimension Reduction* process is more efficient [10, 14].

As described in figure 6, this process reduces the input matrix to a smaller square matrix of l by l. It also gives two projection matrices which can bring the rows and columns of this smaller matrix back to the bases of the original input. It is a randomised procedure to get a ϵ-rank approximation [16] and, therefore, the error will be at the order of the size of the largest truncated singular value. Our *Jupyter Notebook* provides more details [3].

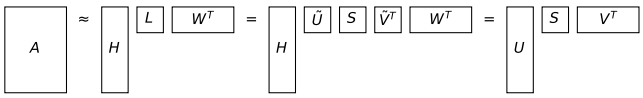

Figure 6: Approximate SVD via dimension reduction

The Randomised Dimension Reduction process can also be applied to the CCA or MCA analysis of two spatial-temporal fields. Similar to the QR Product SVD, it has the advantage that the SVD is applied to a small l by l matrix, as seen in figure 7.

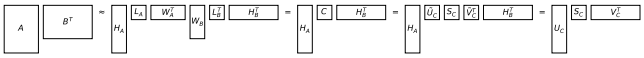

Figure 7: Approximate SVD for product of two fields

To see the effect of dimension reduction on such a matrix product, let's generate various Gaussian Random Fields and compare their cross-correlation matrix with a reduced version. Figure 8 shows

that the results are terrible for fields with a small α, but high autocorrelation allows for optimisation without aquiring much error. Again, we performed our analysis with two generated fields which

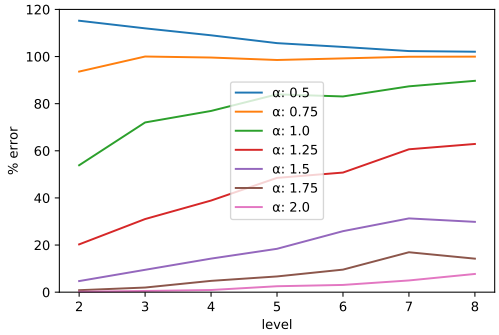

Figure 8: Error after SVD of approximated fields

correlated highly. Whether this correlation influences the amount of error after dimension reduction is left for further research.

Note that, unlike the coarsening procedure, the reduction is not applied on each time-slice of the spatial field, but rather on the spatially-flattened time-series. Therefore, the level of spatial autocorrelation may not be as important as the level of temporal autocorrelation. The proper analysis of this is left for further research.

The reduction of the number of dimensions of each input dataset is actually advised by some researchers, as a method to filter out noise [1]. Especially when the number of temporal samples is small, outliers and random fluctuations could affect the result [4]. 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. In a CCA, where the fields are standardized, this resulting mode may appear important even though it stems from noise. In an MCA, the variance of the noise will be low, so the chance that it will appear as an important mode is less [4]. One method of finding the right level of filtering is by bootstrapping/cross-validating the results [15].

3. APPLICATIONS

Phenology is the science that studies the timings of recurring biological events such as leafing and blooming as well as their causes and variations in space and time. The Extended Spring Indices (SI-x) are a suite of models that transform daily temperatures into consistent phenological metrics [20]. In this section, we use a new long-term (1989 to 2014) and high spatial resolution (1km) version of the Bloom index, which was recently generated for the US by adapting the SI-x models to a cloud computing environment [12].

Time series of remotely sensed images can be used to derive various land surface phenological metrics. One of these metrics is the so-called Start of Season (SOS), which indicates the beginning of photosynthetic activity in plants. Here we use a SOS product specifically made for the US by processing time series of the Advanced Very High Resolution Radiometer (AVHRR) sensor [18].

Calculated over a square subsection of the USA, we found the Bloom field to have $\alpha \approx 1.4$ and the SOS field to have $\alpha \approx 0.6$ [3].

3.1 Approximate SVD via Spatial Coarsening

After coarsening of a spatial subsection of the Bloom field for 1989, we found an error of a few percent, in agreement with the line for $\alpha = 1.5$ on figure 4. For SOS, the error was slightly higher, as expected [3]. After coarsening both fields for all time periods, we performed an MCA. The error was substantial, around 50%, but in agreement with the results of generated fields plotted in figure 5.

3.2 Approx. SVD via Dimension Reduction

After dimension reduction of a spatial subsection of the Bloom field for 1989, we found a negligible error. This could be expected, as explained in the article by Halko et al. [10]. For SOS, the error was higher but still small [3]. After reducing both fields for all time periods, we performed an MCA. The error was substantial, between 30% and 60% depending on specific parameters. These results are in agreement with those of generated fields plotted in figure 8.

4. FURTHER QUESTIONS

It may be interesting to extend this research to fields other than the Gaussian Random Field. This type was chosen because its structure scale can be captured in a single parameter α. In many applications, however, the dataset does not resemble such a Gaussian Random Field.

Additionally, it would be an improvement to relax the assumption that the auto-correlation in the time direction is similar to that in the spatial directions. In fact, it may even be more realistic to have different levels of autocorrelation in the x and in the y direction.

Can similar tricks be used to the generalised MCA/CCA analysis, where the input to the SVD in a concatenation of multiple cross-correlation matrices [5, 13]?

Can the dimension reduction be applied to the spatial part of the spatial-temporal fields, before it is flattened?

Can a randomised algorithm be developed to estimate Moran's I or some other measure of autocorrelation, using a sample of the data?

5. REFERENCES

[1] T. P. Barnett and R. Preisendorfer. Origins and levels of monthly and seasonal forecast skill for us surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, 115(9):1825–1850, 1987.

[2] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

[3] L. Bogaardt. Information loss during size reduction depending on structure scale. `https://github.com/phenology/`, 2018.

[4] C. S. Bretherton, C. Smith, and J. M. Wallace. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, 5(6):541–560, 1992.

[5] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, pages 283–319, 1970.

[6] T. F. Chan. An improved algorithm for computing the singular value decomposition. *ACM Trans. Math. Softw.*, pages 72–83, 1982.

[7] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, pages 211–218, 1936.

[8] G. Eshel. *Spatiotemporal Data Analysis*. Princeton University Press, 2011.

[9] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970.

[10] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[11] L. J. Hubert, R. G. Golledge, and C. M. Costanzo. Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 13(3):224–233, 1981.

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

[13] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, pages 433âĂŞ–451, 1971.

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

[15] R. E. Livezey and T. M. Smith. Covariability of aspects of north american climate with global sea surface temperatures on interannual to interdecadal timescales. *Journal of Climate*, 12(1):289–302, 1999.

[16] P.-G. Martinsson. Randomized methods for matrix computations and analysis of high dimensional data. *ArXiv*, 2016.

[17] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[18] B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen. Measuring phenological variability from satellite imagery. *Journal of Vegetation Science*, 5:703–714, 1994.

[19] S. Rey. Pysal. `http://pysal.readthedocs.io`, 2009–2013.

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

[21] M. Tygert. Suggested in personal communication, 10 2017.

[22] H. von Storch and F. W. Zwiers. *Statistical Analysis In Climate Research*. Cambridge University Press, 1999.