Detailed dataset descriptions

Five high dimensional datasets have been chosen for analysis. The choices are based on public availability, an attempt to cover different data analysis domains and an emphasis on a high number of dimensions. Each dataset was standardized as a pre-processing step.

Faces in the Wild

We use a subset of the Faces in the Wild[1] dataset [1] modified to grey scale images of 40×40 pixels. The dimensions of the dataset are $N = 1000$ and $D = 1600$ and the signal-to-noise ratio was estimated to be $S_1 = 388$. This is a dataset based on news photographs which have been cropped to contain only faces and background. These faces are captured "in the wild" which means there is a lot of variation in both background and foreground, which fits well with the generative model for probabilistic PCA.

When the intensity in different pixel locations vary together across images, that gives rise to a covariance structure between pixels. If we take two pixels close to the centre of the images (i.e. probably within the face region) then these two pixels will often vary together across images, for example due to skin colour or the location of face features such as eyes. This correlation structure gives rise to the principal directions in PCA, which can be referred to as eigenfaces [2].

Missing data processes affect image and video archive restoration, where defects are seen due to blothces (dirt or gelatine sparkle) and scratches on film documents [3]. In other situations missing can affect images as artefacts such as logos, or subtitles that may have been added at some point, but are undesirable at a later point where the original is no longer available.

MNIST

Here we use the test set part of the MNIST[2] [4] database for handwritten digits. The dimensions of the dataset are $N = 10000$ and $D = 784$ and the signal-to-noise ratio was estimated to be $S_1 = 51$. The dataset contains grey scale images of handwritten digits of 28×28 pixels. The digits have been centered which means that most pixel variation is seen in the center part of the image while very little variation is seen in along the edges and corners of the images. This does not fit too well with the generative model for probabilistic PCA, expecting similar noise variance in all dimensions.

As for the face images when the intensity in different pixel locations vary together across images this gives rise to a covariance structure between pixels. The shape of the ten digits 0-9 and the different styles in people's handwriting determine which pixels display some kind of covariance and give rise to the principal directions in PCA.

NCI60

[1] `http://tamaraberg.com/faceDataset/`

[2] `http://yann.lecun.com/exdb/mnist/`

This is the NCI60 Cancer Microarray Project[3][4] [5]. The dimensions of the dataset are $N = 64$ and $D = 6830$ and the signal-to-noise ratio was estimated to be $S_1 = 874$.
This dataset contains observations from cDNA microarrays used to explore patterns of gene expression in different cell lines, related to cancer screening. The 64 cell lines which are derived from human tumours of different origin and the expression of 6830 different genes in these cell lines constitutes the dataset.
Different genes having similar gene expression levels across cell lines gives rise to a covariance structure between genes. This covariance structure determines the principal directions in the PCA, which can be referred to as eigengenes [6].
Experiments with microarrays are frequently affected by missing values and some of the possible reasons for missing data are insufficient resolution, image corruption and dust or scratches on the slide, spotting problems, poor hybridization and fabrication errors [7, 8]. Systematic reasons for missing data may occur due to the robotic methods used to create them [7].

Food pairing
This is the food pairing dataset[5] [9], modified to a bag-of-words style dataset containing $N = 56498$ recipes and a total of $D = 381$ ingredients and the signal-to-noise ratio was estimated to be $S_1 = 5.5$. The data was originally used to investigate general patterns in ingredient combinations across recipes and regional cuisines.
The covariance structure in this dataset arise from similar uses of ingredients across recipes, eliciting a covariance between certain ingredients. This may be due to deserts where most of the recipes using sugar, butter or flour, or some set of basic ingredients that are often used in Thai cuisine for example.

FashionMNIST
Here we use the test set part of the Zalando Fashion[6] dataset [10]. This is an MNIST-like dataset using grey scale images of clothing items instead of handwritten digits. The dimensions of the dataset are $N = 10000$ and $D = 784$ and the signal-to-noise ratio was estimated to be $S_i = 221.7$. The dataset has been introduced to the machine learning community as a replacement or alternative to the much used MNIST data set. Here the covariance structure is related to the similarities in the appearance of shoes, dresses, jeans and shirts.

References

[1] Tamara L Berg, Alexander C Berg, Jaety Edwards, and David A Forsyth. Who's in the picture. In *Advances in neural information processing systems*, pages 137–144,

[3] `http://genome-www.stanford.edu/nci60/`
[4] `https://web.stanford.edu/~hastie/ElemStatLearn/`
[5] `https://www.nature.com/articles/srep00196`
[6] `https://github.com/zalandoresearch/fashion-mnist`

2005.

[2] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[3] Raphaël Bornard, Emmanuelle Lecan, Louis Laborelli, and Jean-Hugues Chenot. Missing data correction in still images and image sequences. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 355–361. ACM, 2002.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Douglas T Ross, Uwe Scherf, Michael B Eisen, Charles M Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S Jeffrey, Matt Van de Rijn, Mark Waltham, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics*, 24(3):227, 2000.

[6] Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[7] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[8] Chia-Chun Chiu, Shih-Yao Chan, Chung-Ching Wang, and Wei-Sheng Wu. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC systems biology*, 7(6):S12, 2013.

[9] Yong-Yeol Ahn, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1, 2011.

[10] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.