Formalising the subjective interestingness of a linear projection of a data set: two examples

[This is a 'work-in-progress' paper]

Tijl De Bie
University of Bristol, Intelligent Systems Laboratory
Bristol, United Kingdom
tijl.debie@gmail.com

ABSTRACT

The generic framework for formalising the subjective interestingness of patterns presented in [2] has already been applied to a number of data mining problems, including itemset (tile) mining [3, 8, 9], multi-relational pattern mining [18, 19, 20], clustering [10], and bi-clustering [12, 11]. Also, it has been pointed out without providing detail that also Principal Component Analysis (PCA) [7] can be derived from this framework [2]. This short note describes work-in-progress aiming to show in greater detail how this can be done. It also shows how the framework leads to a robust variant of PCA when used to formalise the subjective interestingness of a data projection for a user who expects outliers to be present in the data.

Categories and Subject Descriptors

H.4 [**Information systems applications**]: Data mining

General Terms

Theory, Algorithms

Keywords

Principal Component Analysis, Robust PCA, subjective interestingness

1. INTRODUCTION

This short note gives two examples of how the framework from [2], can be used to formalise the subjective interestingness of a linear projection of a data set. Thus it illustrates how the framework can lead to different approaches for linear dimensionality reduction depending on the prior beliefs of the user, illustrating the importance of this initial interaction with the user.

We consider two types of prior beliefs in particular. The first one of these leads to an algorithm identical to Principal Component Analysis (PCA). The second one, which is suited for users who feel they have no accurate belief about the spread of the data but only about the order of magnitude of that spread, can be thought of as a robust (outlier insensitive) alternative to PCA that appears to be novel.

This note sweeps all details under the carpet, and leaves a number of important questions unanswered. These details and questions will be resolved in a later publication. The hope is that this short note further demonstrates the usefulness of the framework from [2] across the breadth of exploratory data mining research. It helps in elucidating when a certain pattern is interesting to a given user, depending on the beliefs of that user.

In this particular study, it shows that PCA is not the best approach for users who anticipate the presence of outliers. While this will come as no surprise to many, this is a formal and rigorous demonstration of why that is the case, and additionally offers an alternative method that is appropriate when outliers are expected by a user.

2. SUBJECTIVE INTERESTINGNESS IN A NUTSHELL

2.1 Notation

Scalars are denoted with standard face, vectors with bold face lower case, and matrices with bold face upper case letters. The i'th data point is denoted as $\mathbf{x}_i \in \mathbb{R}^d$ with d the dimensionality of the data space. The matrix containing all data points transposed $\mathbf{x}_i'$ $(i = 1, \ldots, n)$ as its rows is denoted as $\mathbf{X} \in \mathbb{R}^{n \times d}$.

2.2 Projection patterns

In the general framework of [2], we formalised patterns as any property the data satisfies. In this paper, the particular kind of pattern considered can formalised as a constraint on the data of the form:

$$\mathbf{X}\mathbf{w} = \mathbf{p},$$

where $\mathbf{w} \in \mathbb{R}^d$, referred to as a weight vector (also known as the loadings), has unit norm and parameterises the pattern. The vector $\mathbf{p} \in \mathbb{R}^n$ specifies the value of the projections of the data points onto the weight vector $\mathbf{w}$. The fact that the projections of all data points onto a given weight vector $\mathbf{w}$ are equal to specific values is clearly a property a data set may or may not have, and revealing it to a user provides clear information to that user restricting the set of possible values the data set can have.

Although ideally any possible $\mathbf{w} \in \mathbb{R}^d$ can be considered, in practice only a finite though large number of them can be considered due to the lack of finite code for the set of real numbers. Similarly, the values of $\mathbf{p}$ cannot be specified to an infinite accuracy. This short note brushes over these issues, which can be dealt with rigorously by assuming they are specified up to a certain accuracy. A rigorous treatment of these issues is deferred to a later publication.

2.3 The subjective interestingness of projection patterns

In [2], the interestingness of a pattern (defined generically as any constraint on the value of the data) is formalised as the trade-off between the description length of the pattern, and its subjective information content. More specifically, the *subjective interestingness* of a pattern is formalised as its subjective information content divided by its description length. Here we very briefly summarize this framework, and start outlining how it can be applied to the kind of patterns of interest in this paper, namely projection patterns.

It is reasonable to consider the *description length* as constant, independent of $\mathbf{w}$ and $\mathbf{p}$. Indeed, this amounts to assuming that each possible $\mathbf{w}$ requires the same description length, and that $\mathbf{p}$ is shown with constant absolute precision. The latter is the case when e.g. the projections are visualized on a computer screen or printed on paper. If the values of $\mathbf{p}$ are normalised before visualizing, then the description length is not exactly constant as also the normalising factor needs to be specified, which requires a variable length code if the normalisation factor is unbounded. However, in practice this should always account for a very small part of the description length of the pattern.

The *subjective information content* is minus the logarithm of the probability that the pattern is present, where the probability is computed with respect to the so-called background distribution, which represents the belief state of the user about the data.

The belief state can be modelled assuming a certain set of prior beliefs (expressed as constraints on the expected values of certain test statistics given the background distribution). Among all distributions satisfying these constraints, the *background distribution* is the one with maximum entropy.

Each time a pattern is revealed to the user, the user's background distribution changes. More specifically, it is conditioned on the presence of the pattern just revealed.

3. INTERESTING PROJECTIONS WHEN NO OUTLIERS ARE EXPECTED

3.1 Prior beliefs and the background distribution

A user not expecting any outliers will be able to express an expectation about the value of the average two-norm squared of the data points:

$$\mathbb{E}_{\mathbf{X}\sim P}\left\{\frac{1}{n}\sum_{i}^{n}\mathbf{x}_i'\mathbf{x}_i\right\} \quad = \quad \sigma^2.$$

To determine the value for σ user involvement appears to be inevitable at first sight. However, below it will become clear that the ordering of projection patterns according to interestingness is in fact independent of the value of σ, so in practice the exact value will not need to be known.

It is well known (and easy to derive) that the distribution of maximum entropy given this prior belief constraint on the scatter matrix of the data points is a product distribution of multi-variate normal distributions with mean $\mathbf{0}$ and covariance matrix $\sigma\mathbf{I}$. I.e. the density function for each of the data points $\mathbf{x}$ is:

$$p(\mathbf{x}) \quad = \quad \frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2\sigma^2}\right).$$

Thus, the product of n such distributions, one for each of the data points, is the background distribution formalising a user's prior belief state about the data set, when that user does not anticipate the presence of outliers.

3.2 The subjective interestingness of a projection pattern

It is well-known that the probability distribution of an orthogonal transformation of a normal random variable is again a normal random variable, with the same mean and with a covariance matrix that is transformed accordingly. In the current context, with $\mathbf{W}$ an orthogonal matrix (i.e. $\mathbf{W}'\mathbf{W} = \mathbf{W}\mathbf{W}' = \mathbf{I}$), and with $\mathbf{z} = \mathbf{W}'\mathbf{x}$, it holds that:

$$\begin{aligned} p(\mathbf{z}) \quad &= \quad \frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{\mathbf{z}'\mathbf{z}}{2\sigma^2}\right), \\ &= \quad \prod_{k=1}^{d}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z_k^2}{2\sigma^2}\right). \end{aligned}$$

I.e., the distribution of $\mathbf{z}$ is a product distribution with a factor for each of the components of $\mathbf{z}$. Thus, the marginal distribution for the first component, z_1, is given by:

$$p(z_1) \quad = \quad \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z_1^2}{2\sigma^2}\right).$$

Referring to the first column of $\mathbf{W}$ as $\mathbf{w}$ (and note that $\mathbf{w}'\mathbf{w} = 1$ follows from $\mathbf{W}'\mathbf{W} = \mathbf{I}$), this means that the projections $\mathbf{X}\mathbf{w} = \mathbf{p}$ of all data points follow this normal distribution, and thus the subjective information content of a projection pattern specified by this equality is equal to:

$$\begin{aligned} &\text{SubjectiveInformationContent}\left(\mathbf{X}\mathbf{w} = \mathbf{p}\right) \\ = \quad &-\log\left(p(\mathbf{X}\mathbf{w} = \mathbf{p})\right), \\ = \quad &\frac{n}{2}\log(2\pi) + \frac{1}{2\sigma^2}\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w}. \end{aligned}$$

As the descriptional complexity is constant, this is proportional to the subjective interestingness.

3.3 The maximiser of the interestingness is the maximiser of the variance

PCA's goal is to maximise $\mathbf{w}'\mathbf{X}'\mathbf{X}'\mathbf{w}$ subject to the constraint $\mathbf{w}'\mathbf{w} = 1$, which is clearly equivalent with maximising this subjective interestingness. PCA can thus be regarded as finding the projection pattern with maximal subjective interestingness for the user not expecting any outliers.

3.4 Subsequent iterations

After revealing the first projection pattern, the background distribution is conditioned on the fact that $\mathbf{X}\mathbf{w} = \mathbf{p}$. The updated background distribution is then a product distribution of multivariate standard normal distributions on the subspace orthogonal to $\mathbf{w}$. The result of that is that the subjective information of patterns in subsequent iterations is computed as for the first pattern after deflating the data: considering only the component of the data points orthogonal to $\mathbf{w}$. This is precisely the way PCA works.

4. INTERESTING PROJECTIONS WHEN OUTLIERS ARE EXPECTED

With a slightly different prior belief that assumes the presence of outliers (leading to a heavy-tailed background distribution), a method that can be thought of as a robust version of PCA is obtained.

4.1 Prior beliefs and the background distribution

As prior beliefs, now the following is used:

$$\mathbb{E}_{\mathbf{X}\sim P}\left\{\frac{1}{n}\sum_{i}^{n}\log\left(1+\frac{1}{\rho}\mathbf{x}_i'\mathbf{x}_i\right)\right\} \quad = \quad c.$$

This kind of prior belief specifies an expectation on a measure of the spread of the data, which amplifies contributions from points with small norm relative to the data points with large norm through a log transformation. Thus, using such a prior belief rather than say a prior belief on the second moment considers outliers in the data relatively more probable. The smaller the value of ρ, the less important the constant term in the argument of the logarithm will be, the more logarithmic this statistic will therefore vary with the norm of $\mathbf{x}_i$, and thus the more tolerant this model will be to outliers. Informally: rather than determining an expectation on the spread of the data, for small values of ρ it determines an expectation on the *order of magnitude* of the spread of the data.

For convenience in the following derivations, let us introduce the function

$$\kappa(\nu) \quad = \quad \psi\left(\frac{\nu+d}{2}\right)-\psi\left(\frac{\nu}{2}\right),$$

where ψ represents the digamma function. In the sequel the value of $\kappa^{-1}(c)$ will need to be used, denoted as ν for brevity. Then, the initial background distribution can be derived by relying on [22], where it is shown that the maximum entropy distribution subject to the specified prior information is the product of independent multivariate standard t-distributions with density function p defined as:

$$p(\mathbf{x}) \quad = \quad \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\sqrt{(\pi\rho)^d}\Gamma\left(\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{1}{\rho}\mathbf{x}'\mathbf{x}\right)^{\frac{\nu+d}{2}}},$$

with one factor in this product distribution for each data point. Here Γ represents the gamma function.

Note that for $\rho,\nu\to\infty, \frac{\rho}{\nu}\to\sigma^2$ this tends to the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. For $\rho=\nu=1$ this is a multivariate standard Cauchy distribution, which is so heavy-tailed that its mean is undefined and its second moment is infinitely large. Thus, this type of prior beliefs can clearly model the expectation of outliers to varying degrees.

4.2 The subjective interestingness of a projection pattern

To compute the subjective information content, note that the density function for the transformed variable $\mathbf{z}=\mathbf{W}'\mathbf{x}$ with $\mathbf{W}$ an orthogonal matrix is given as:

$$p(\mathbf{z}) \quad = \quad \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\sqrt{(\pi\rho)^d}\Gamma\left(\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{1}{\rho}\mathbf{z}'\mathbf{z}\right)^{\frac{\nu+d}{2}}}.$$

Now, the density function for the marginal distribution of a t-distribution with given covariance matrix is again a t-distribution density with the same number of degrees of freedom, obtained by simply selecting the relevant part of the covariance matrix [13, 15]. With $\mathbf{w}$ denoting the first column of $\mathbf{W}$, this means that the density function for $z_1=\mathbf{w}'\mathbf{x}$, the first component of $\mathbf{z}$, is:

$$p(z_1) \quad = \quad \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\rho}\Gamma\left(\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{1}{\rho}z_1^2\right)^{\frac{\nu+1}{2}}}.$$

Written in terms of $\mathbf{x}$, this is:

$$p(\mathbf{x}'\mathbf{w}) \quad = \quad \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\rho}\Gamma\left(\frac{\nu}{2}\right)}\cdot\frac{1}{\left(1+\frac{1}{\rho}\mathbf{w}'\mathbf{x}\mathbf{x}'\mathbf{w}\right)^{\frac{\nu+1}{2}}}.$$

Thus, the subjective information content of a pattern stating that $\mathbf{X}\mathbf{w}=\mathbf{p}$ is:

$$\begin{aligned}&\text{SubjectiveInformationContent}\left(\mathbf{X}\mathbf{w}=\mathbf{p}\right)\\ =\quad &\frac{\nu+1}{2}\sum_{i=1}^{n}\log\left(1+\frac{1}{\rho}(\mathbf{x}_i'\mathbf{w})^2\right)+\text{a constant}.\end{aligned}$$

Again, as the description length is constant, this is proportional to the subjective interestingness.

4.3 Maximising the interestingness using a robust version of PCA

Taking into account that $\mathbf{w}'\mathbf{w}=1$ (as required in the patterns considered and as imposed by the orthogonality of $\mathbf{W}$), maximising the subjective interestingness is thus equivalent to solving the following problem:

$$\begin{aligned}\max_{\mathbf{w}} \quad &\sum_{i=1}^{n}\log\left(\rho+(\mathbf{x}_i'\mathbf{w})^2\right),\\ \text{s.t.} \quad &\mathbf{w}'\mathbf{w}=1.\end{aligned}$$

The method of Lagrange multipliers leads to the following optimality condition for the subjective information content:

$$\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\rho+(\mathbf{x}_i'\mathbf{w})^2}\right)\mathbf{w} \quad = \quad \lambda\mathbf{w}.$$

Note that the matrix on the left hand side is propotional to essentially a weighted empirical covariance matrix for the data, where points contribute more if they have a smaller value for $(\mathbf{x}_i'\mathbf{w})^2$: the weight for $\mathbf{x}_i\mathbf{x}_i'$ is $\frac{1}{\rho+(\mathbf{x}_i'\mathbf{w})^2}$.

Although this optimisation problem is not convex and the optimality conditions do not admit a closed form solution in terms of e.g. an eigenvalue problem, a modified version of the power method for solving eigenvalue problems empirically appears to be a good heuristic approach. The algorithm goes as follows:

1. Solve the eigenvalue problem $\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)\mathbf{w}=\lambda\mathbf{w}$ for the dominant eigenvector,[1] further denoted $\mathbf{w}^{(0)}$. This vector is normalised to unit norm.

[1]This amounts to solving the problem for $\rho\to\infty$, which is essentially equivalent to PCA. This is no coincidence as for $\rho,\nu\to\infty, \frac{\rho}{\nu}\to\sigma^2$ the background distribution is an isotropic multivariate Gaussian distribution, as noted above.

2. Iterate from $k = 1$ until convergence or maximum number of iterations reached:
   (a) $\mathbf{v}^{(k)} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i'}{\rho + (\mathbf{x}_i' \mathbf{w}^{(k-1)})^2}\right) \mathbf{w}^{(k-1)}$.
   (b) $\mathbf{w}^{(k)} = \frac{\mathbf{v}^{(k)}}{\|\mathbf{v}^{(k)}\|}$.

Clearly this is not guaranteed to converge to the global optimum, but in practice it appears to perform well. Whether it always converges to a local optimum is left as an open question in this note.

The effect of the parameter ρ is as follows. For a smaller value of ρ, the tail of the background distribution can be heavier, as then the nonlinearity of the logarithm in the prior belief constraint will affect data points of smaller magnitude. The effect of this is that outliers (for which $(\mathbf{x}_i'\mathbf{w})^2$ may be very large) will not weigh in as strongly as they would in PCA, as the contribution of $\mathbf{x}_i\mathbf{x}_i'$ to what can be thought of as a reweighted covariance matrix is reduced, and relatively more so than for data points for which $(\mathbf{x}_i'\mathbf{w})^2$ is small as compared to ρ (for which the reduction is roughly constant). Informally speaking, ρ is a soft threshold on the squared distance along $\mathbf{w}$ beyond which data points will no longer be able to bias the solution in their own direction.

Interestingly, just like in PCA where the value of σ has no effect on which pattern is most interesting, here the value of ν and thus of c has no effect on which projection is the most interesting one. (Though σ and c do affect the value of the interestingness in both cases.) This significantly reduces the demands on the user in specifying their prior beliefs.

4.4 Subsequent iterations

A property of the multivariate t-distribution is that the conditional distribution conditioned on the value of any of the dimensions is again a multivariate t-distribution, though with a different number of degrees of freedom and a different covariance matrix [15]. Thus, after revealing the values of the projections $\mathbf{p}$, the updated background distribution is again a multivariate t-distribution for the parts of the data points orthogonal to $\mathbf{w}$ from the first pattern. The next pattern can be found essentially by projecting the data points onto the orthogonal complement of $\mathbf{w}$ and repeating the same procedure.

5. EXPERIMENT

To illustrate the robustness of the PCA alternative derived in the previous section, consider a dataset consisting of 1000 data points sampled from a Gaussian distribution with mean $\mathbf{0}$ and with covariance matrix $\begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$, to which a further 100 'outliers' are added, sampled from a Gaussian distribution with mean $\mathbf{0}$ and with covariance matrix $\begin{pmatrix} 16 & 12 \\ 12 & 13 \end{pmatrix}$.

The weight vector resulting from standard PCA is shown with a full red line in Fig. 1. The black dash-dotted lines show the weight vectors retrieved by the robust PCA method described, with values for ρ equal to $1, 10$, and 100. The largest value of these resulted in the line closest to the PCA result. The green dashed line shows the weight vector that would have been found using standard PCA had there been no outliers at all (i.e. computed just on the first 1000 data points).

The left figure shows the resulting weight vectors on top of a scatter plot of all data points, clearly showing that the PCA result is determined primarily by the outliers. The right figure shows the same resulting weight vectors on top of a scatter plot of only the first 1000 data points (excluding the outliers). Clearly, the robust PCA version is much less strongly affected by the outliers and primarily determined by the dominant variance direction in the bulk of the data points excluding the outliers.

6. DISCUSSION AND FURTHER WORK

This note shows how PCA can be derived as an instantiation of the framework from [2] for deriving subjective interestingness of exploratory data mining patterns. Additionally, it shows how prior beliefs reflecting the expectation that outliers may be present in the data lead to an alternative to PCA that is less sensitive to such outliers.

Robust PCA is an important research topic that has been studied for decades, see e.g. [1, 14, 6, 21] for a few recent references. Often the problem is tackled as an instance of projection pursuit (and also our algorithm could be viewed as such) [4, 5], by making use of a robust estimator of the covariance matrix [17, 16], or by making additional assumptions about the nature of the interesting aspects of the data and the corrupting noise process. The algorithm derived in this note appears to be most strongly related to the algorithm from [14], but further study into connections between the two is required.

In further work we will enhance the rigour of the derivations, attempt to establish the convergence of the algorithm for the robust version of PCA, and investigate the utility of other alternatives to PCA that are useful for other relevant kinds of prior belief states. E.g. it is relatively straightforward to add assumptions on anisotropy of the data to the prior beliefs in both the derivation of PCA and of the robust version of PCA, as well as assumptions about the expected average of the data points not being the origin. However also altogether different kinds of prior beliefs could be of interest.

Acknowledgements

This work is supported by the ERC Consolidator Grant FORSIED.

7. REFERENCES

[1] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[2] T. De Bie. An information-theoretic framework for data mining. In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[3] T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3):407–446, 2011.

[4] Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. 1973.

[5] Peter J Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

[6] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

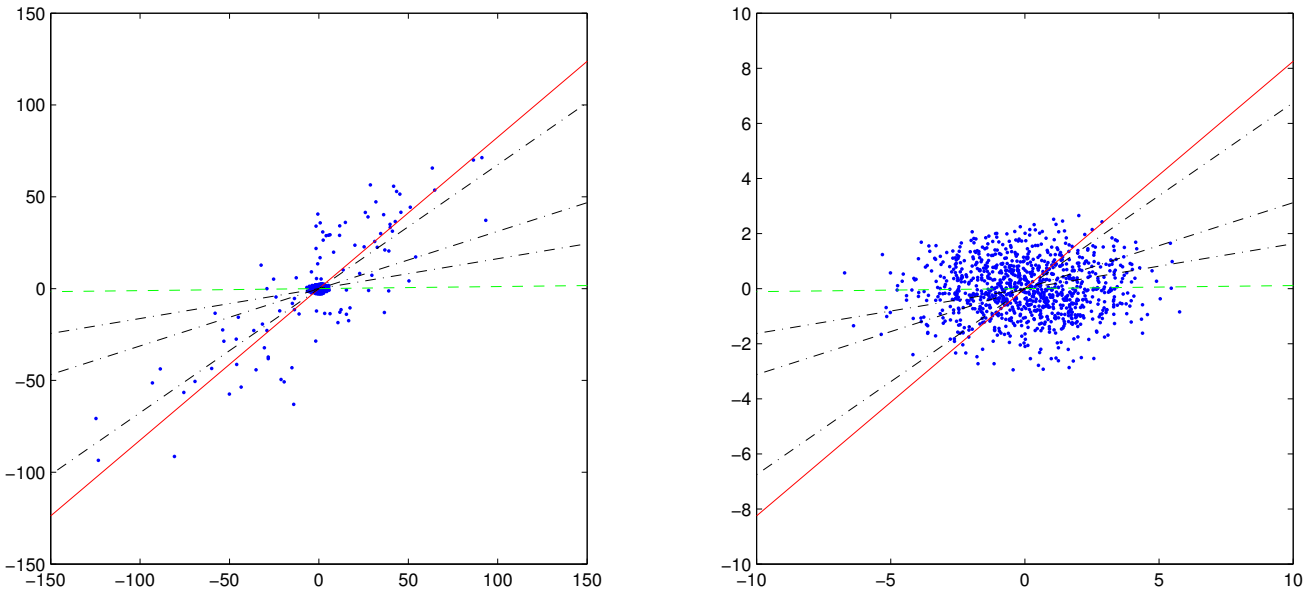

Figure 1: The left plot shows a scatter plot of all data points including outliers, with weight vectors of standard PCA (continuous red line), as well as the robust PCA with values for $\rho = 1, 10, 100$ (black dash-dotted line) and standard PCA on the data points excluding the 100 outliers (green dashed line). The right plots shows the same results but now without visualising the outliers in the scatter plot.

[7] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[8] K.-N. Kontonasios and T. De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proc. of the 2010 SIAM International Conference on Data Mining (SDM)*, 2010.

[9] K.-N. Kontonasios and T. De Bie. Formalizing complex prior information to quantify subjective interestingness of frequent pattern sets. In *Proc. of the 11th International Symposium on Intelligent Data Analysis (IDA)*, 2012.

[10] K.-N. Kontonasios and T. De Bie. Subjectively interesting alternative clusterings. *Machine Learning*, 2013.

[11] K.-N. Kontonasios, J. Vreeken, and T. De Bie. Maximum entropy modelling for assessing results on real-valued data. In *Proc. of the IEEE International Conference on Data Mining (ICDM)*, 2011.

[12] K.-N. Kontonasios, J. Vreeken, and T. De Bie. Maximum entropy models for iteratively identifying subjectively interesting structure in real-valued data. In *Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery from Databases (ECML-PKDD)*, 2013.

[13] S. Kotz and S. Nadarajah. *Multivariate t distributions and their applications*. Cambridge University Press, 2004.

[14] Yongmin Li, L-Q Xu, Jason Morphett, and Richard Jacobs. An integrated algorithm of incremental and robust pca. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I–245. IEEE, 2003.

[15] Michael Roth. On the multivariate t distribution. Technical Report LiTH-ISY-R-3059, Department of Electrical Engineering, Linköping universitet, April 2013.

[16] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[17] Peter J Rousseeuw and Bert C Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, 1990.

[18] E. Spyropoulou and T. De Bie. Interesting multi-relational patterns. In *Proc. of the IEEE International Conference on Data Mining (ICDM)*, 2011.

[19] E. Spyropoulou, T. De Bie, and M. Boley. Interesting pattern mining in multi-relational data. *Data Mining and Knowledge Discovery*, 2013.

[20] E. Spyropoulou, T. De Bie, and M. Boley. Mining interesting patterns in multi-relational data with n-ary relationships. In *Discovery Science (DS)*, 2013.

[21] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[22] K. Zografos. On maximum entropy characterization of pearson's type II and VII multivariate distributions. *Journal of Multivariate Analysis*, 71(1):67–75, 1999.