

Sparsity in the context of learning from high dimensional data

Bob Durrant and Ata Kabán - University of Birmingham

ICARN International Workshop - Liverpool 26th September 2008

Nearest Neighbour Theorem (Beyer et. al. [3])

- Let F_m , $m \in \mathbb{N}$ be an infinite sequence of data distributions.
- Let $\{\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)}, \dots, \mathbf{x}_N^{(m)}\}$ be a random sample of N independent data points distributed as F_m , and let $\mathbf{x}^{(m)}$ be an arbitrary data point drawn from F_m .
- For each m , let $\|\cdot\| : \text{dom}(F_m) \rightarrow \mathbb{R}^+$ be a function that takes a point from the domain of F_m and returns a positive real number.
- Let $p > 0$ denote an arbitrary positive constant, and assume that both $\mathbb{E} [\|\mathbf{x}^{(m)}\|^p]$ and $\text{Var} [\|\mathbf{x}^{(m)}\|^p]$ are finite with $\mathbb{E} [\|\mathbf{x}^{(m)}\|^p] \neq 0$.

Nearest Neighbour Theorem (Beyer et. al. [3])

If:

$$\lim_{m \rightarrow \infty} \frac{\text{Var} [\|\mathbf{x}^{(m)}\|_p]}{\text{E} [\|\mathbf{x}^{(m)}\|_p]^2} = 0$$

then $\forall \varepsilon > 0$:

$$\lim_{m \rightarrow \infty} \Pr \left[\max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| \leq (1 + \varepsilon) \cdot \left(\min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| \right) \right] = 1$$

Where the operators $\text{E}[\cdot]$ and $\text{Var}[\cdot]$ refer to the theoretical expectation and variance of the distributions comprising F_m , and the probability is over a random sample of size N drawn from F_m .

Converse Near Neighbour theorem (Kabán and Durrant)

With the same notation as before, assume the sample size N is large enough for $\mathbb{E} [\|\mathbf{x}^{(m)}\|^p] \in \left[\min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|^p, \max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|^p \right]$ to hold.

If $\forall \varepsilon > 0$,

$$(1) \quad \lim_{m \rightarrow \infty} \Pr \left[\max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| \leq (1 + \varepsilon) \cdot \left(\min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| \right) \right] = 1$$

then

$$(2) \quad \lim_{m \rightarrow \infty} \frac{\text{Var} [\|\mathbf{x}^{(m)}\|^p]}{\mathbb{E} [\|\mathbf{x}^{(m)}\|^p]^2} = 0$$

Proof (Slide 1)

Write $DMIN_m = \min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|$ and $DMAX_m = \max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|$.

Without loss of generality we can take $DMIN_m \neq 0$, so that rewriting the condition (1) we get:

$$(3) \quad \lim_{m \rightarrow \infty} \Pr [DMAX_m \leq (1 + \varepsilon) \cdot DMIN_m] = 1 \quad \Rightarrow$$

$$(4) \quad \lim_{m \rightarrow \infty} \Pr \left[\frac{DMAX_m}{DMIN_m} \leq (1 + \varepsilon) \right] = 1 \quad \Rightarrow$$

$$(5) \quad \lim_{m \rightarrow \infty} \frac{DMAX_m}{DMIN_m} = 1$$

From the definition of convergence in probability.

Proof (Slide 2)

Then using Slutsky's Theorem twice on (5) we have both that:

$$(6) \quad \lim_{m \rightarrow \infty} \frac{DMAX_m^p}{DMIN_m^p} = 1, \text{ and also: } \lim_{m \rightarrow \infty} \frac{DMIN_m^p}{DMAX_m^p} = 1$$

Now, by our initial assumption that $E[\|\mathbf{x}^{(m)}\|^p] \in [DMIN_m^p, DMAX_m^p]$, and as the power function $(\cdot)^p$ is monotonic increasing on \mathbb{R}^+ , we can see that:

$$(7) \quad \frac{DMIN_m^p}{DMAX_m^p} \leq \frac{\|\mathbf{x}_j^{(m)}\|^p}{E[\|\mathbf{x}^{(m)}\|^p]} \leq \frac{DMAX_m^p}{DMIN_m^p}, \quad \forall j \in \{1, 2, \dots, N\}$$

Proof (Slide 3)

Then, by the squeeze rule, it follows that:

$$(8) \quad \lim_{m \rightarrow \infty}^P \left\{ \frac{\|\mathbf{x}^{(m)}\|_p}{\mathbb{E} [\|\mathbf{x}^{(m)}\|_p]} \right\} = 1$$

Which is a sequence of random variables converging in probability to a constant.

Finally, since convergence in probability implies convergence in distribution, we can conclude that the associated sequence of variances converges to zero, that is:

$$\lim_{m \rightarrow \infty} \text{Var} \left[\frac{\|\mathbf{x}^{(m)}\|_p}{\mathbb{E} [\|\mathbf{x}^{(m)}\|_p]} \right] = \lim_{m \rightarrow \infty} \frac{\text{Var} [\|\mathbf{x}^{(m)}\|_p]}{\mathbb{E} [\|\mathbf{x}^{(m)}\|_p]^2} = 0$$

as required. ■

Is it possible for $RV_m \rightarrow 0$?

Consider the relative variance $RV_m = \frac{\text{Var}[\|\mathbf{x}^{(m)}\|^p]}{\mathbb{E}[\|\mathbf{x}^{(m)}\|^p]^2}$ and the distance function $\|\cdot\|^p$, interpreted as a p -norm over some metric space. Using definitions and making no new assumptions about the data structure, for a random vector $\mathbf{x}^{(m)} = (x_1, x_2, \dots, x_m)^T$ of arity m we can rewrite RV_m as:

$$RV_m = \frac{\text{Var}[\sum_{i=1}^m |x_i|^p]}{\mathbb{E}[\sum_{i=1}^m |x_i|^p]^2} = \frac{\sum_{i=1}^m \sum_{j=1}^m \text{Cov}[|x_i|^p, |x_j|^p]}{\sum_{i=1}^m \sum_{j=1}^m \mathbb{E}[|x_i|^p] \mathbb{E}[|x_j|^p]}$$

Then, if the numerator grows no slower than the denominator with increasing m the RV_m will not tend to zero, and the distances between the data points will remain separated.

Linear Latent Variable Models

Consider an L -dimensional linear latent variable model, where the observed dimensions x_i are modelled by linear combinations of the L latent variables y_l ($l \in \{1, 2, \dots, L\}$, $L < \infty$), plus additive noise. In such a case the model is:

$$x_i = \sum_{l=1}^L a_{il} y_l + \delta_i, \quad \forall i \in \{1, 2, \dots, m\}$$

We assume that the noise term is zero mean iid and independent of the systematic factors y_l ($\delta \sim N(0, \sigma)$, say), s.t. all of the structure of the data can be expressed by a combination of latent variables.

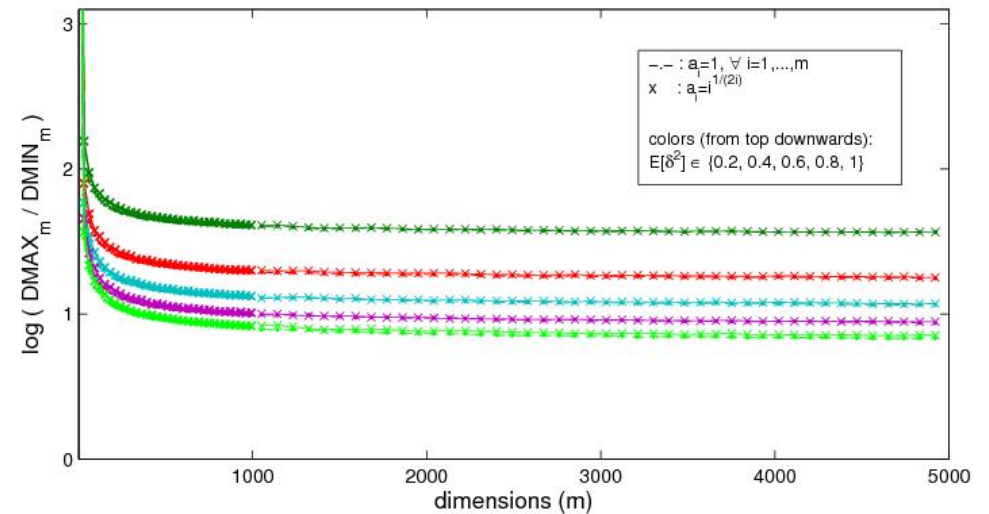
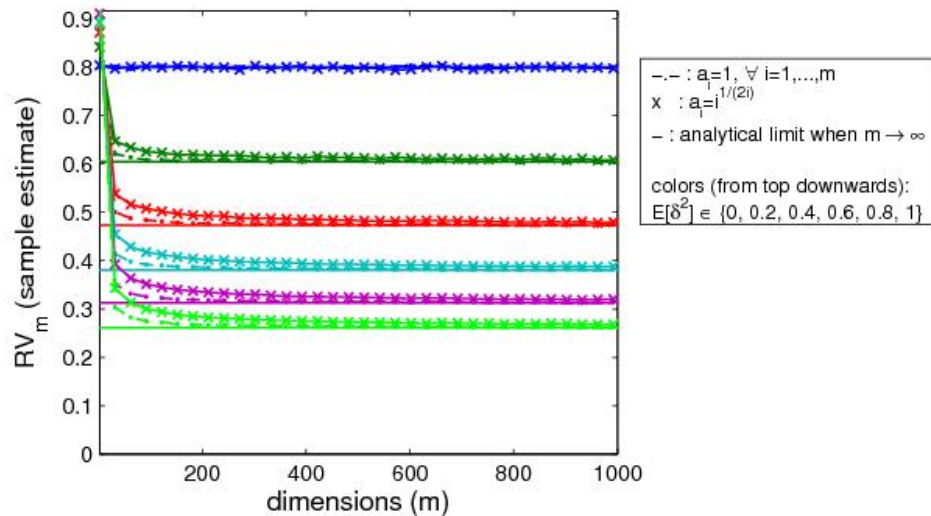
We also assume that $\text{Var}(y_l^2) \neq 0$.

It is possible to show that for such a model the Euclidean distance does not concentrate in the noiseless case, nor in the presence of noise provided that the contribution from the systematic components grows no slower than that from the noise.

Likewise, provided that the noise is bounded away from zero, then if the systematic contribution grows no slower than the dimensionality m the Euclidean distance does not concentrate for these models.

Numerical Validation

Example showing RV_m and $\log[DMAX_m/DMIN_m]$ for increasing m .
 $L = 1, y \sim Uniform[0, 2], \delta \sim N(0, \sigma^2), \sigma \in [0, 1]$. (σ varies).
 The a_i were designed such that $\lim_{m \rightarrow \infty} a_m^2 = 1$.

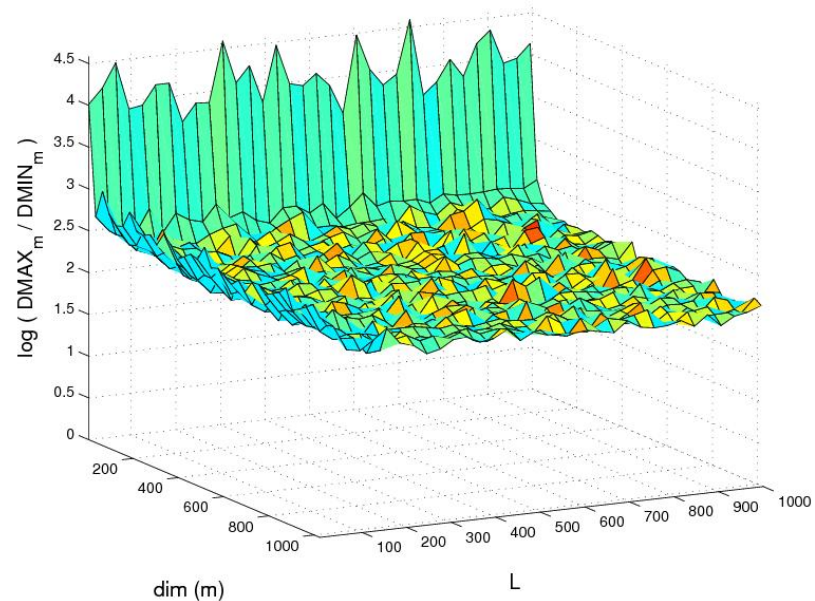
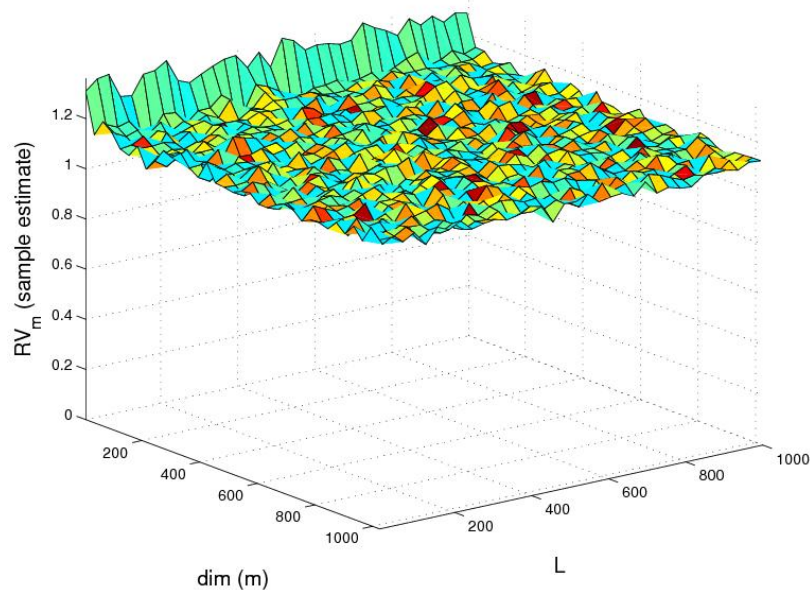


Empirical estimates are superimposed on the corresponding analytic limits, and we see the sequence of RV_m estimates converge is in agreement with these limits.
 (Results averaged from 10 repeats over 15,000 points for each m .)

Numerical Validation

Example showing increasing latent dimensionality L , alongside increasing data dimensionality m .

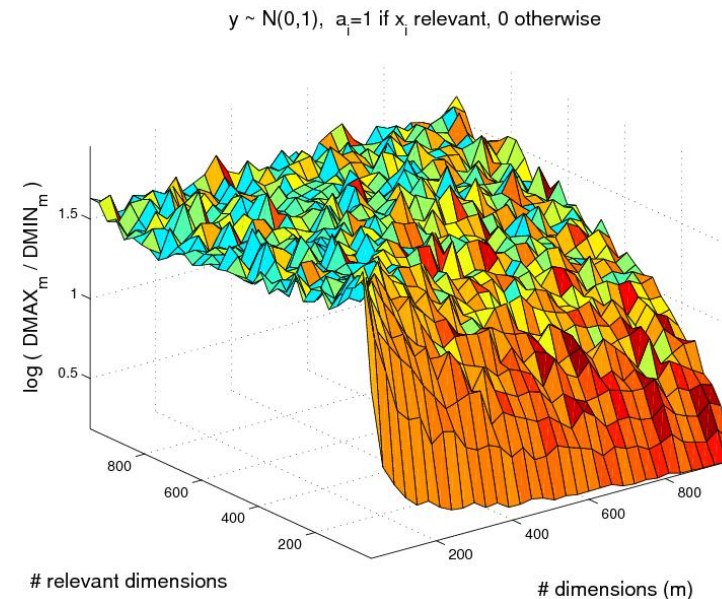
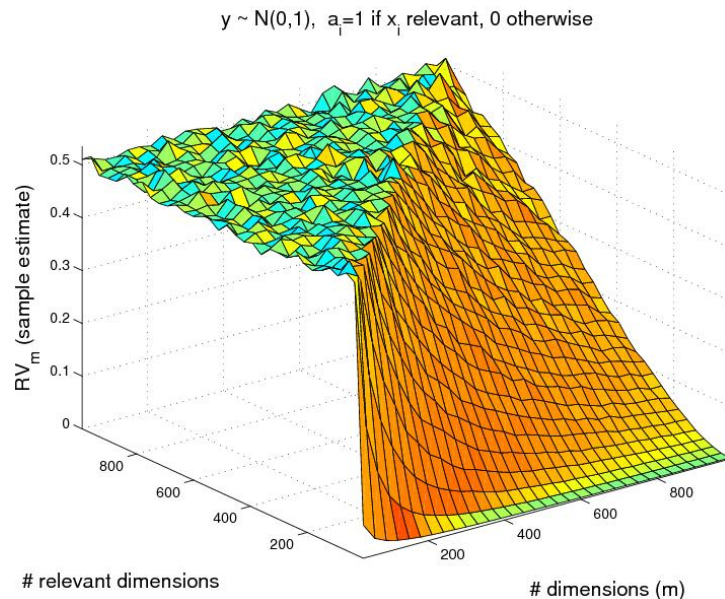
Here $y_i \sim N(0, 1)$, $\delta \sim N(0, 1)$, $a_i \in [0, 3]$ and the pairs $\mathbf{a}_l, \mathbf{a}_k$ are not orthogonal. Each estimate is based on 20,000 points from the model.



Effect of 'irrelevant' dimensions

Consider again the case where $L = 1$. We call a dimension *irrelevant* if its weight a_i^2 is zero, i.e. its only contribution is its noise term.

We have seen that to avoid concentration, the cumulative contribution from the systematic factors must grow no slower than the data dimensionality. Here we show the effect of varying the number of irrelevant dimensions in a regression model $L = 1, y \sim N(0, 1)$ and $\delta \sim N(0, 1)$.



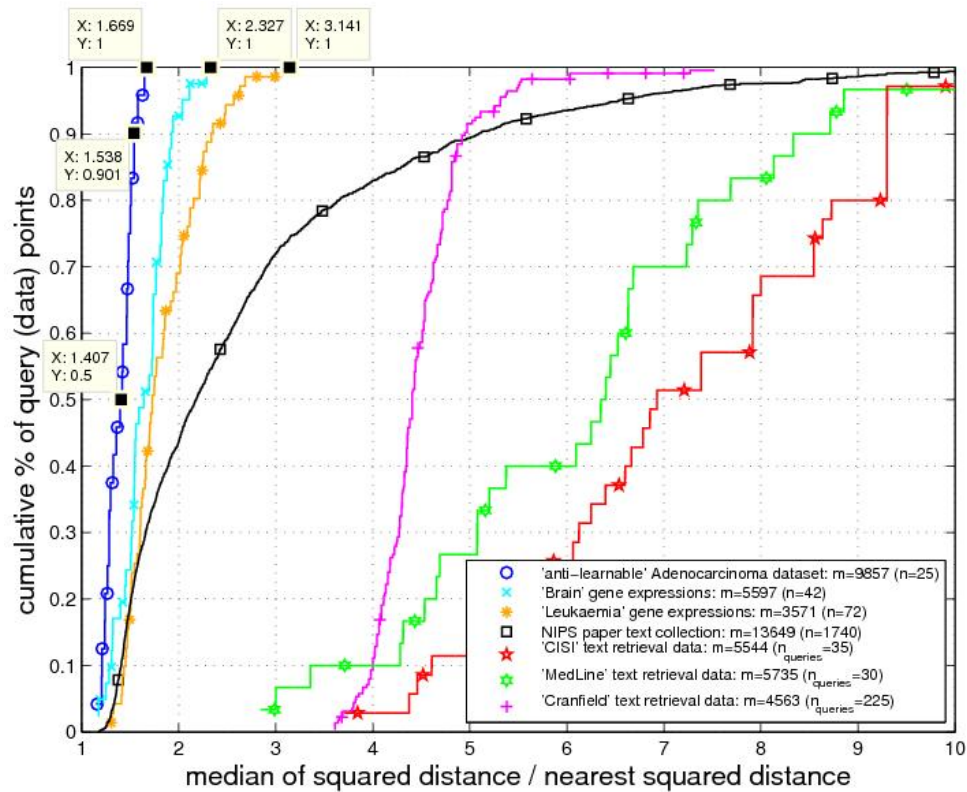
Examining distance concentration in real data sets

We compared sets of gene expression data ($3000 < m < 10000$) to text data sets with comparable dimensionality.

- Text data were selected because, although such sets can exhibit extremely high dimensionality ($m = |Dictionary|$), the data are well-known to have a rich correlation structure and there have been many successes reported in this problem domain [7].
- Gene expression data were selected because, apart from their practical importance, the number of relevant dimensions in such sets can be extremely low compared to the dimensionality and some of these data succeed in breaking the best classifiers [4],[8].

Examining distance concentration in real data sets

A comparison of real data sets demonstrates that the gene expression data is much more concentrated than the text data:



Conclusions

- We established the converse of Beyer et al's theorem, formulating the necessary conditions for the distance concentration phenomenon.
- By examining a broad class of non-iid data models, linear latent variable models, we identified settings in which the Euclidean distance does not concentrate under reasonable conditions. (This complements earlier work in the field, where the focus has been on non-Euclidean metrics or iid dimensions).
- The class of models we considered has wide usage, so our analysis gives some practically valuable guidance and explanation as to when and why distance concentration will be a problem (or not) in a HD data setting.

Some observations regarding ICA and Projection Pursuit

Consider again the single latent variable case. Under the assumption of iid zero-mean unit variance latent variables y_i , we have that, up to an additive $o(m)$ term:

$$(9) \quad RV_m = \frac{\text{Var}[y^2]}{\left(\mathbb{E}[y^2] + \frac{\sum_{i=1}^m \mathbb{E}[\delta_i^2]}{\|\mathbf{a}_1\|^2} \right)^2}$$

$$(10) \quad = \frac{\mathbb{E}[y^4]/\mathbb{E}[y^2] - 1}{\left(1 + \frac{\sum_{i=1}^m \mathbb{E}[\delta_i^2]}{\|\mathbf{a}_1\|^2} \right)^2}$$

$$(11) \quad = \frac{\text{kurt}(y) + 2}{\left(1 + \frac{\sum_{i=1}^m \mathbb{E}[\delta_i^2]}{\|\mathbf{a}_1\|^2} \right)^2}$$

We want to maximise RV_m , and this can therefore be achieved by maximising either of $\text{kurt}(y)$ or the SNR: $\frac{\|\mathbf{a}_1\|^2}{\sum_{i=1}^m \mathbb{E}[\delta_i^2]}$.

Some observations regarding ICA and Projection Pursuit

Likewise it is possible to show in the many latent variables case ($\infty > L > 1$) that maximising RV_m , modulo $o(m)$ terms, is equivalent to maximising:

$$RV_m = \frac{\sum_{l=1}^L \text{kurt}(y_l) + L(L + 2)}{\left(L + \frac{\sum_{i=1}^m (\delta_i^2)}{\|\mathbf{a}_1\|^2}\right)^2}$$

Our result therefore appears to indicate one reason why maximising kurtosis as an objective function leads to good results in ICA and PP for signal denoising; by using such an approach we are maximising the relative variance RV_m of the distance between data points, and therefore the separation of the sources y_l .

References

- [1] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. Proc. Int. Conf. Database Theory, pp. 420-434, 2001.
- [2] C.C. Aggarwal and P.S. Yu. The IGrid Index: Reversing the dimensionality curse for similarity indexing in high dimensional space. Proc. ACM Conf. KDD, pp.119-129, 2000.
- [3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? Proc. Int. Conf. Database Theory, pp. 217-235, 1999.
- [4] R. Clarke, H. W. Resson, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer, vol. 8, pp. 37-49, Jan. 2008.
- [5] B. Everitt. An introduction to latent variable models. Chapman and Hall, 1984.
- [6] D François, V Wertz, and M Verleysen. The concentration of fractional distances. IEEE Trans. on Knowledge and Data Engineering, vol 19, no 7, July 2007.
- [7] T. Joachims: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. ECML 1998: 137-142.
- [8] A. Kowalczyk. Classification of anti-learnable biological and synthetic data, Proc. PKDD, pp. 176-187, 2007.
- [9] S.T. Roweis, and Z. Ghahramani: A Unifying Review of Linear Gaussian Models. Neural Computation 11(2): 305-345, 1999.