Maxvol for Machine Learning

Philip Blagoveschensky
Mirfarid Musavian
Maria Sindeeva
Ivan Golovatskikh

December 20, 2018

Introduction

The goal is to select square or rectangular submatrix $\tilde{A} \in R^{k\times m}$ of matrix $A \in \mathbb{R}^{n\times m}$ such that:

$$vol(\tilde{A}) = \begin{cases} |\det \tilde{A}|, & \text{if } k = m \\ \sqrt{\det(\tilde{A}^* \tilde{A})}, & \text{if } k > m \end{cases} \qquad \longrightarrow \qquad \underset{\tilde{A}}{\text{maximum.}}$$

Square case

Submatrix $\tilde{A} \in \mathbb{R}^{m \times m}$ of a matrix $A \in \mathbb{R}^{n \times m}$ is called dominant if a swap of any single row of $\tilde{A}$ with a row of A, not already presented in $\tilde{A}$, does not increase the volume of $\tilde{A}$.

- *Maxvol*[1] looks for the dominant submatrix.
- Can be computed with $\Theta(cnm)$ complexity, where c is the number of iterations.

[1] Goreinov S. A. et al. How to find a good submatrix

Rectangular case

Skoltech
Skolkovo Institute of Science and Technology

Let $A \in \mathbb{R}^{n\times m}$ be a full column rank matrix.

- *Rectangular maxvol*[2] can be used to find a dominant submatrix $\tilde{A} \in \mathbb{R}^{k\times m}$, i.e. a submatrix with large volume.
- Complexity is $\Theta(nm^2)$.

[2] Mikhalev A., Oseledets I. V. Rectangular maximum-volume submatrices and their applications

Skoltech
Skolkovo Institute of Science and Technology

- Recommendation systems, the "cold start" problem[3].
- Functions approximation.
- In the case of linear regression $y = \tilde{A}x + \theta$, maximizing such objective function leads to minimizing noise variance:

$$\mathsf{Var}(x) = (\tilde{A}^* \tilde{A})^{-1} \sigma^2.$$

(*D-optimality* criterion[4])

[3] Liu N. N. et al. Wisdom of the better few: cold start recommendation via representative based rating elicitation

[4] J. Kiefer, Optimum experimental designs V, with applications to systematic and rotatable designs

Our hypotheses

Skoltech
Skolkovo Institute of Science and Technology

- Maxvol performs feature selection well when number of features is much greater than number of samples
    - With assumption that there are some core ('true') features, and the rest are their transformations (linear and non-linear)
    - With assumption that there are some core ('true') features, and the rest are noise
- Maxvol performs sample selection well by select most informative samples of dataset

Experiments with synthetic datasets

Skoltech
Skolkovo Institute of Science and Technology

Test whether *Maxvol* selects features well. Generate synthetic datasets for this purpose by sampling a probability distribution.

- Two classes – B and R. Define events B and R to be "the object is of class B" and "the object is of class R" respectively. $P(B) = 0.75, P(R) = 0.25$.
- Objects have k "true features" $z_1, \ldots, z_k$. Define $\sigma_{\min}, \sigma_{\max}$ – range of standard deviations for true features. For each $i \in \{1, \ldots, k\}$ if class is B then feature z_i has distribution $z_i \sim N(2, \sigma_i)$; if class is R then feature z_i has distribution $z_i \sim N(4, \sigma_i)$. For all i it holds that $\sigma_i \in [\sigma_{\min}, \sigma_{\max}]$.
- If we add redundant features, will *maxvol* discard them?

Synthetic: linearly dependent features

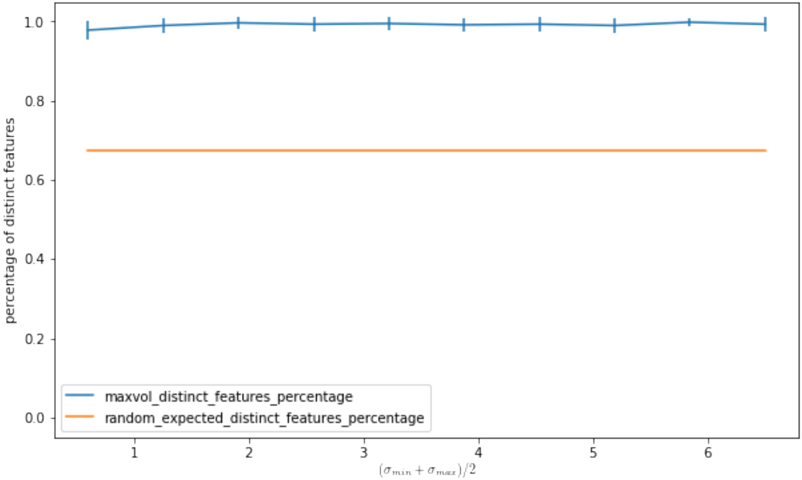
- $P(B) = 0.75$
- k normally distributed "true features" $z_1, \ldots, z_k$ with means 2 and 4 for B and R respectively.
- For each $i \in \{1, \ldots, k\}$ introduce 5 features $x_i^{(1)}, \ldots, x_i^{(5)}$ which are true features plus small noise, i.e. $\forall j \in \{1, \ldots, 5\}\, x_i^{(j)} = z_i + \eta_i^{(j)}$ with $\eta_i^{(j)} \sim N(0, 0.1)$.
- Generate dataset $A \in \mathbb{R}^{n \times 5k}$ with $n \geq 5k$ where each row is a sample from this distribution. Column number $5j + i$ represents feature $x_i^{(j)}$. Hence each row has 5 copies of each true feature value (plus small noise).

Synthetic: linearly dependent features

- $A \in \mathbb{R}^{n \times 5k}$. Each row has 5 copies of each true feature value (plus small noise).
- We normalize each feature (column of A), choose k random samples (rows of A) and run *maxvol* on that to select k features (columns).
- The hypothesis is that it will select very few duplicated features (features made from the same "true feature"), because determinant should be very small if two almost linearly dependent columns are selected.

Synthetic: linearly dependent features

$k = 60$ "true features", $\sigma_{\max} - \sigma_{\min} = 1$. Error bars represent values no more than two empirical stds from empirical mean.



Synthetic: transformations by continuous functions

For each index $i \in \{1, \ldots, k\}$ of "true feature" introduce 5 features $x_i^{(1)}, \ldots, x_i^{(5)}$ which are generated by a continuous function applied to the "true feature" plus small noise and a constant, i.e. $\forall j \in \{1, \ldots, 5\}$ $x_i^{(j)} = f_j(z_i + c_i^{(j)} + \eta_i^{(j)})$ with $\eta_i^{(j)} \sim N(0, 0.1)$.

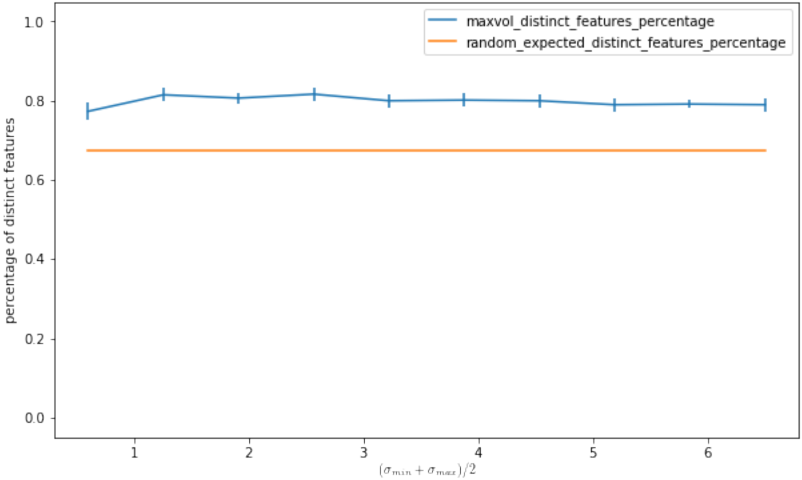
1. $f_1(z) = z$
2. $f_2(z) = e^z$
3. $f_3(z) = \sqrt{(|z|)}$
4. $f_4(z) = z^2$
5. $f_5(z) = z^3$

Synthetic: transformations by continuous functions

- We generate dataset $A \in \mathbb{R}^{n \times 5k}$ with $n \geq 5k$ where each row is a sample from this distribution. Column number $5j + i$ represents feature $x_i^{(j)}$. Hence each row has 5 transformations of each true feature value.
- We normalize each feature (column of A), choose k random samples (rows of A) and run *maxvol* on that to select k features (columns).

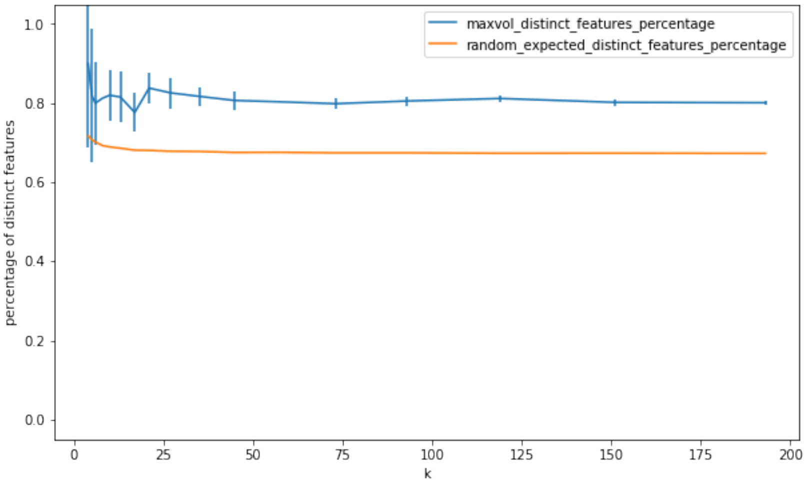
Synthetic: transformations by continuous functions

$k = 60$ "true features", $\sigma_{\max} - \sigma_{\min} = 1$. Error bars represent values no more than two empirical stds from empirical mean.



Synthetic: transformations by continuous functions

$\sigma_{\min} = 3, \sigma_{\max} = 4$, k changing. Error bars represent values no more than two empirical stds from empirical mean.



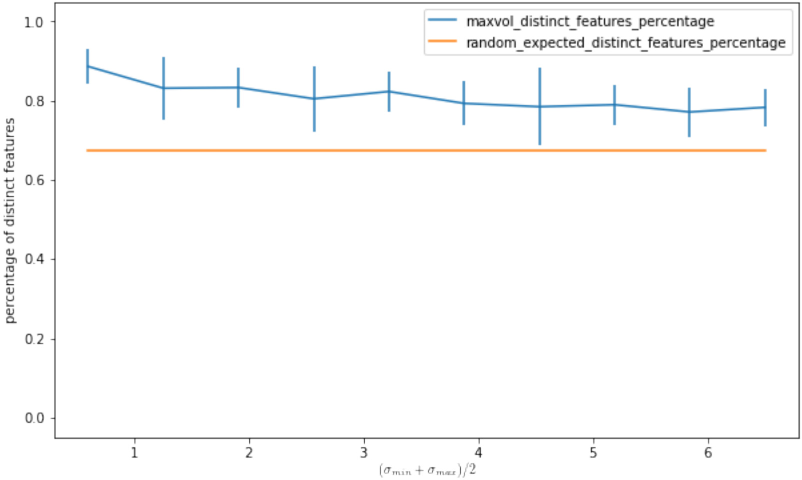
What if only one class?

Does *maxvol* select distinct features because we have two classes which have different means of distributions of features?
Let's check by setting $P(B) = 1, P(R) = 0$.

Single class, continuous transformations

Skoltech
Skolkovo Institute of Science and Technology

$k = 60$ "true features", $\sigma_{\max} - \sigma_{\min} = 1$. Error bars represent values no more than two empirical stds from empirical mean.



Experiments on real-world data
ARCENE

Skoltech
Skolkovo Institute of Science and Technology

Dataset:

- Mass-spectrometric data from healthy and cancer patients: each feature indicates the abundance of proteins in human sera having a certain mass value
- 10 000 features in total: 7 000 real sensor data, 3 000 distractor features with no predictive power
- Data for only 100 patients
- Full-rank matrix

`rect_maxvol` applied:

- Directly to the dataset
- To the scaled dataset

Evaluated using accuracy of perceptron trained on the transformed dataset.

Experiments on real-world data
ARCENE

In both scaled and unscaled settings the results of 4 models are compared for several numbers of features:

- Trained on `rect_maxvol`-selected features
- Trained on a randomly selected subset of features (accuracy is averaged over several random feature choices)
- Trained on all features
- Trained on PCA transformation of the data

Experiments on real-world data

ARCENE

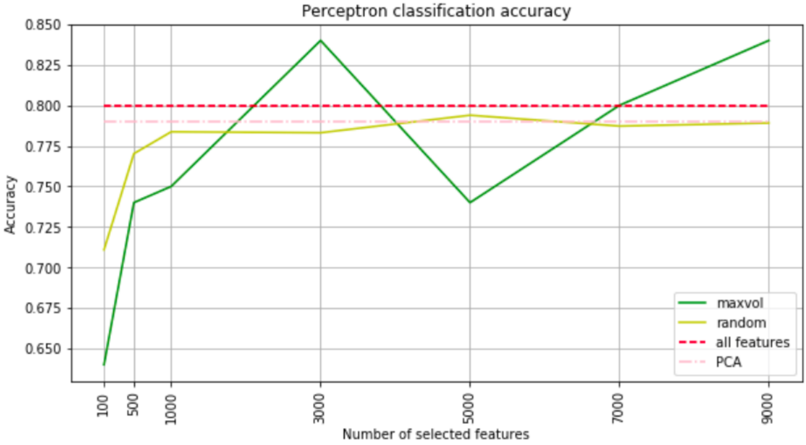


Figure: No scaling

Experiments on real-world data
ARCENE

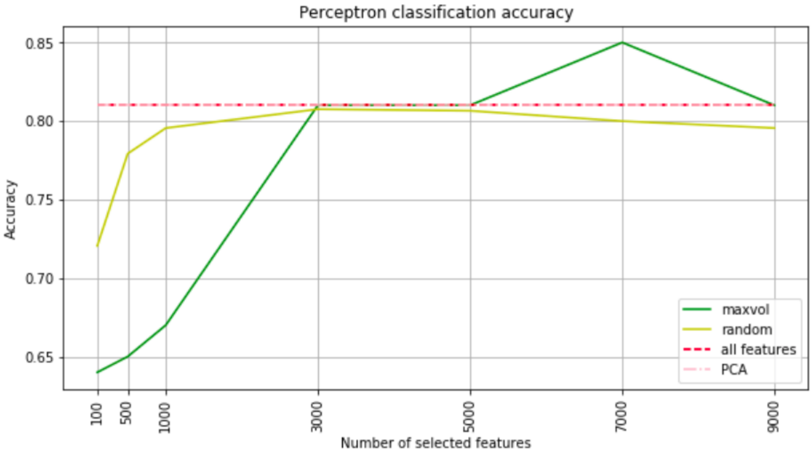


Figure: MinMax scaling

Experiments on real-world data

ARCENE

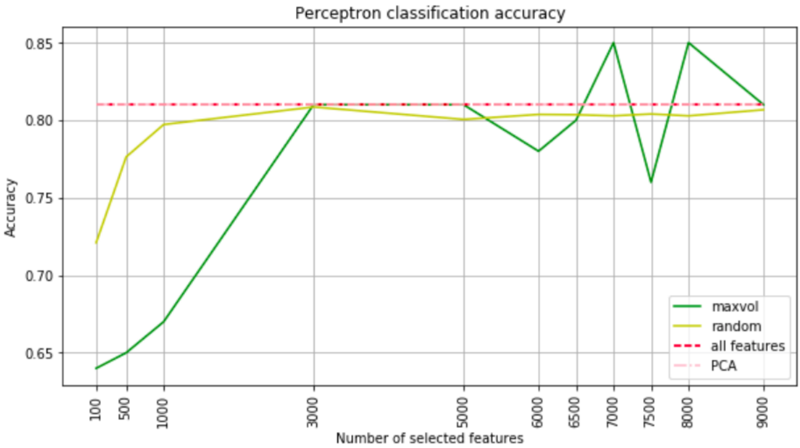


Figure: MinMax scaling in more detail

Experiments on real-world data

MNIST

Dataset:

- Dataset of images of handwritten digits
- 784 features: 28×28-pixel images
- 60 000 images for training
- Rank-deficient matrix

`rect_maxvol` applied:

- To select features
- To select samples

Evaluated using accuracy of perceptron trained on the transformed dataset.

Dataset matrix is column rank deficient: unable to apply `rect_maxvol` directly to the dataset.
We tried two main ideas on this dataset:

- Select the most representative features (a way of dimensionality reduction)
- Select constrained amount of representative examples to speed up training.

Skoltech
Skolkovo Institute of Science and Technology

Feature selection strategy (based on paper[5]):

- Compute the rank-k SVD approximation to the dataset Y,

$$Y_k = U\Sigma V^T$$

- Apply `rect_maxvol` to V and select corresponding features in the original dataset as representative ones.

Subsampling strategy: peform the same steps, but use `rect_maxvol` on U, instead of V

[5] Liu N. N. et al. Wisdom of the better few: cold start recommendation via representative based rating elicitation

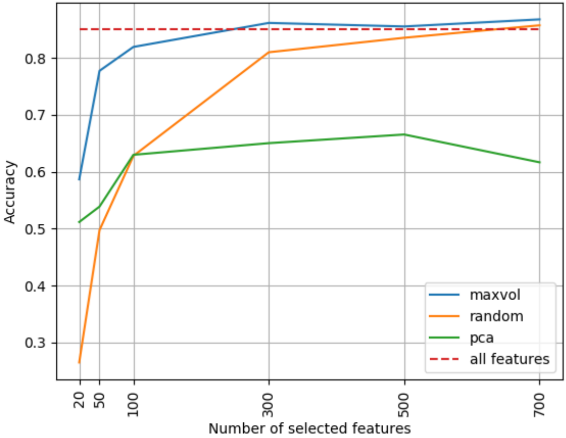


Figure: Feature selection

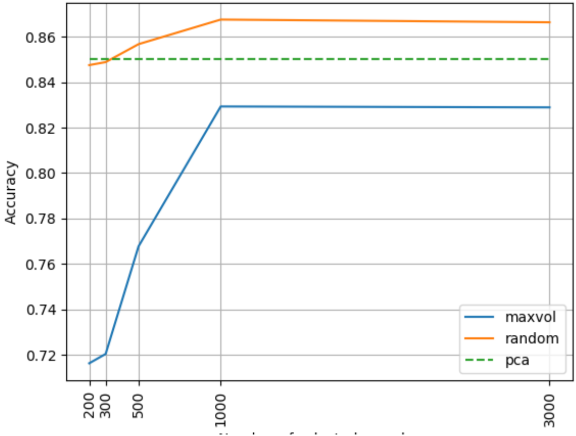


Figure: Sampling

Skolkovo Institute of Science and Technology

Thank you!