UDC 514.742.2:004.658.2

A VECTOR ALGEBRA PERSPECTIVE ON DATA DIMENSIONALITY REDUCTION

Harbar A.E., student

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Kniaziuk N.V. – PhD in Physics and Mathematics, Associate Professor

Abstract. The paper studies methods of applying vector algebra to optimize machine learning tasks. Using the Human Activity Recognition dataset, a mathematically rigorous primary data analysis is conducted: the Gram matrix and correlation structure are examined, positive semi-definiteness of the correlation matrix is proved, and multicollinearity is quantified via the condition number $k(\mathrm{X^TX}) \approx 9.56 \cdot 10^4$ and median *VIF* = 22.4. Principal Component Analysis is examined through the lens of singular value decomposition; the orthogonality of principal components is proved. Projection onto $k = 20$ components increases classifier accuracy from 62.8 % to 65.5 %.

Keywords. Vector algebra, Gram matrix, singular value decomposition, principal component analysis, primary data analysis, multicollinearity, condition number, Eckart-Young theorem.

High-dimensional data – datasets in which the number of features p is comparable to or exceeds the number of samples m – pose fundamental mathematical difficulties. The curse of dimensionality, formulated by R. Bellman [1], states that the sample size required for reliable learning grows exponentially: estimating a probability density in a p-dimensional unit cube to accuracy ε requires on the order of ε^{-p} samples. For p = 561, this renders a direct approach practically infeasible.

A second problem is multicollinearity: linear dependence among the columns of the feature matrix leads to ill-conditioning of the normal matrix X^TX and instability of linear model estimators. Both obstacles are overcome by methods grounded in vector algebra.

The objective of this paper is to study methods of applying vector algebra to optimize machine learning tasks: to conduct a mathematically rigorous primary analysis of the high-dimensional Human Activity Recognition (HAR) dataset and to demonstrate how the singular value decomposition of the data matrix underlies Principal Component Analysis (PCA) and provides optimal dimensionality reduction.

The HAR dataset [2] contains readings from three-axis accelerometer and gyroscope of a smartphone sampled at 50 Hz. From raw time series, 561 statistical features were extracted and organized into three physical blocks: time domain (265 features), frequency domain (208 features), and angular characteristics (88 features). The target variable is an activity label from the set {1,…,6}: walking, walking upstairs, walking downstairs, sitting, standing, and lying. The data are represented by the matrix:

$$X \in \mathbb{R}^{m \times p}, \quad m = 10299, p = 561 \tag{1}$$

Notation: $(\|\cdot\|_2)$ – Euclidean L_2 norm of a vector; $(\|\cdot\|_{\mathrm{F}})$ – Frobenius norm of a matrix; $(\widetilde{\mathrm{X}})$ – mean-centred matrix; $(\mathrm{diag}(\cdot))$ – vector of diagonal entries; $(\mathrm{tr}(\cdot))$ – matrix trace; (I_p) – identity matrix of size $(\mathrm{p} \times \mathrm{p})$; all indices start at 1 unless stated otherwise.

Prior to analysis, each *j*-th column of X is standardized: transformed to zero mean and unit variance. For $j = 1,\ldots,p$:

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \tag{2}$$

$$(X_s)_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{3}$$

$$\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \mu_j)^2}, \tag{4}$$

$$C = \frac{X_s^T X_s}{m-1}, \tag{5}$$

where (2) is the sample mean, (3) is the corrected sample standard deviation. Sklearn.StandardScaler [3] is used, which uses population standard deviation. After standardisation the sample covariance matrix (5) has unit diagonal entries, and its trace satisfies:

$$\mathrm{tr}(C) = \sum_{j=1}^{p} \mathrm{Var}\left(X_{s,j}\right). \tag{6}$$

This property was verified computationally: for the HAR dataset tr(C) ≈ 561.054 (the deviation from the exact value *p* is due to finite-sample effects). Standardisation is mandatory for PCA: without it, features with larger scales artificially dominate the decomposition.

The Gram matrix is the matrix of all pairwise inner products of centred feature vectors:

$$G = \tilde{X}^T\tilde{X},\ G_{ij} = \tilde{x}_i \cdot \tilde{x}_j = \sum_{k=1}^{m} \tilde{X}_{ki}\tilde{X}_{kj} \tag{7}$$

The diagonal entry $G_{jj} = \left\|\tilde{x}_j\right\|_2^2$ is the squared norm of the *j*-th feature. The correlation matrix is obtained by normalising the Gram matrix:

$$R = D^{-1/2}GD^{-1/2},\ R_{ij} = \frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}} = \cos\theta_{ij} = \frac{\tilde{x}_i \cdot \tilde{x}_j}{\|\tilde{x}_i\|_2 \cdot \|\tilde{x}_j\|_2} \tag{8}$$

where $D = \mathrm{diag}(G_{11}, \ldots, G_{pp})$. The entry R_{ij} is the cosine of the angle (θ_{ij}) between the centred *i*-th and *j*-th feature vectors in $\mathbb{R}^m$, which coincides exactly with the Pearson correlation coefficient definition. The value $r = +1$ corresponds to $\theta = 0°$ (collinear vectors); $r = 0$ to $\theta = 90°$ (orthogonal).

The matrix *R* is symmetric $(R = R^T)$, has unit diagonal entries $(R_{jj} = 1)$, and is positive semi-definite. For any $v \in \mathbb{R}^p$:

$$v^T R v = \|\tilde{X}D^{-\frac{1}{2}}v\|_2^2 \geq 0. \tag{9}$$

Proof: $v^T R v = v^T D^{-1/2} G D^{-1/2} v = (D^{-1/2}v)^T \tilde{X}^T\tilde{X}(D^{-1/2}v) = \|\tilde{X}D^{-1/2}v\|_2^2 \geq 0$. Consequently, all eigenvalues of *R* are non-negative, and any truncation of the correlation matrix defines a valid covariance structure.

Analysis of *R* revealed a distinct three-block structure of the feature space: the average $|r_{ij}|$ within the time-domain block is 0.306, within the frequency block 0.262, and between blocks only 0.215, confirming within-block redundancy.

The condition number of the normal matrix:

$$k(X^TX) = \frac{\lambda_{\max}(X^TX)}{\lambda_{\min}(X^TX)} = \frac{\sigma_1^2}{\sigma_p^2}. \tag{10}$$

For the HAR dataset: $\sigma_1 = 1206.66,\ \lambda_1 = \sigma_1^2/(m-1) = 141.39,\ k\left(X^TX\right) \approx 9.56 \cdot 10^4$. With such a condition number, the relative error of the OLS estimator $\beta = \left(X^TX\right)^{-1}X^Ty$ may be 10^4 times larger than the relative error in the input data [4].

The Variance Inflation Factor:

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{11}$$

where $R_j^2 - R^2$ regression of x_j on x_k. On a subsample of 30 HAR features: median *VIF* = 22.4, maximum = 80 (threshold = 5). This quantitatively confirms the need for dimensionality reduction.

For any matrix $A \in \mathbb{R}^{m\times p}$ there exists a decomposition (SVD theorem):

$$A = U\Sigma V^T, \quad U^TU = I_p, \quad V^TV = I_p, \tag{12}$$

where $U \in \mathbb{R}^{m\times p}$ contains the left singular vectors; $\Sigma = diag(\sigma_1, \dots, \sigma_p)$, $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$ are the singular values; $V \in \mathbb{R}^{p\times p}$ contains the principal directions (loadings), $V^TV = I_p$. The Frobenius norm is expressed through singular values:

$$\|A\|_F^2 = \sum_{i,j} A_{ij}^2 = \operatorname{tr}(A^TA) = \sum_{j=1}^{p} \sigma_j^2. \tag{13}$$

For HAR: $\|X\|_F^2 = 5{,}777{,}739$ agrees with $\sum\sigma_j^2$ to machine epsilon ($\approx 14 \times 10^{-12}$).

Substituting SVD into the sample covariance matrix $C = X_s^TX_s/(m-1)$:

$$C = V\left(\frac{\Sigma^2}{m-1}\right)V^T = V\Lambda V^T, \quad \lambda_j = \frac{\sigma_j^2}{m-1}. \tag{14}$$

This is the spectral decomposition of C: the columns of V are the eigenvectors and λ_j are the eigenvalues. The fraction of variance explained by the *j*-th component:

$$\rho_j = \frac{\lambda_j}{\sum_p \lambda_k} = \frac{\sigma_j^2}{\|X_s\|_F^2}. \tag{15}$$

For HAR: $\rho_1 = 25.20\%$, $\rho_2 = 8.28\%$, $\rho_1 + \rho_2 = 33.48\%$. The uniform spectrum (small ρ_1) reflects the presence of many significant directions in the data (figure 1).

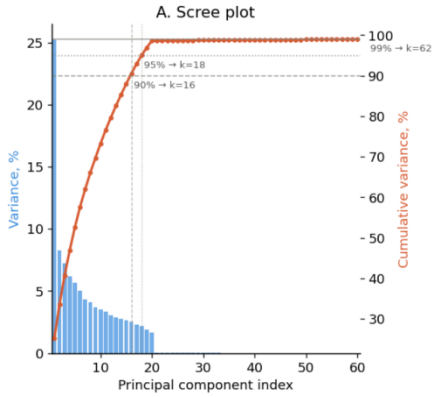

Figure 1 – Explained variance by the first eight principal components (scree plot).

The projection of the data onto the first *k* principal components:

$$Z_k = X_sV_k \in \mathbb{R}^{m\times k}. \tag{16}$$

We prove that the principal components are mutually orthogonal. The covariance matrix of the columns of Z_k:

$$Z^{\mathrm{T}}{}_kZ_k = V^{\mathrm{T}}{}_k(X^{\mathrm{T}}X)V_k = V^{\mathrm{T}}{}_k(V\Sigma^2V^{\mathrm{T}})V_k = \Lambda_k = diag\left(\sigma_1^2, \dots, \sigma_k^2\right). \tag{17}$$

The matrix Λ_k is diagonal because $V^{\mathrm{T}}{}_kV_k = I_k$ (orthonormal columns). This means zero covariances between components: multicollinearity is eliminated by construction. Verification: the maximum off-diagonal entry of $Z^{\mathrm{T}}{}_kZ_k$ for $k = 20$ is $3.7 \cdot 10^{-11}$.

The Eckart–Young theorem [5]: the truncated SVD $X_k = Z_kV^{\mathrm{T}}{}_k$ is the best rank-*k* approximation to X in the Frobenius norm:

$$\|X_s - \hat{X}_k\|_F^2 = \sum_{j>k} \sigma_j^2 = \|X_s\|_F^2 (1 - \rho_1 - \cdots - \rho_k). \quad (18)$$

The relative approximation error decreases with *k*: *k* = 1: 74.8 %; k = 5: 47.4 %; k = 10: 26.7 %; k = 18: 5.0 %; k = 20: 1.4 %.

The effective rank is an entropic measure of intrinsic dimensionality:

$$eff_{rank} = \exp\left(-\sum_{j=1}^{p} \rho_j \ln \rho_j \right). \quad (19)$$

This is the Shannon entropy of the distribution ρ_j. For HAR: $eff_{rank} \approx 16.51$ out of 561. This means that all physically meaningful information about the six activity classes is concentrated in a subspace of dimension ≈17, while the remaining 544 directions carry noise.

To quantify the effect of dimensionality reduction, a logistic regression classifier was trained – a linear model whose decision boundary is a hyperplane in feature space. Data split: 80 % training (8,239 samples), 20 % test (2,060 samples), stratified. Metrics: Accuracy and macro-F1. Results are shown in Table 1.

Table 1 – Classification results for different feature space configurations

Feature Configuration	*p*	Acc., %	F1, %	Dim. reduction / $\sum \rho_j$
Original features	561	62,8	62,7	— / 100 %
PCA, *k* = 10	10	61,1	60,8	−98 % / 73,3 %
PCA, *k* = 18	18	64,8	64,5	−97 % / 95,0 %
PCA, *k* = 20	20	65,5	65,3	−96 % / 98,6 %
PCA, *k* = 30	30	65,1	64,9	−95 % / 98,8 %

The interpretation relies on the prediction error decomposition:

$$\mathbb{E}[(f(x) - y)^2] = \text{Bias}^2(f) + \text{Var}(f) + \sigma_{noise}^2. \quad (20)$$

With all 561 features, the model overfits noise components (small σ_j, j > 20), resulting in high estimator variance. At *k* = 20, removing these components (approximation error 1.4 %) sharply reduces variance. The net result: Accuracy +2.7 pp while reducing feature count 28-fold. The Eckart–Young theorem guarantees the optimality of this choice.

The paper demonstrates that vector algebra tools form a unified mathematical foundation for primary data analysis and optimization of machine learning tasks.

The Gram matrix $G = \widetilde{X}^T \widetilde{X}$ and normalised correlation matrix $R = D^{-1/2} G D^{-1/2}$ enable quantitative characterisation of feature space structure ($R_{ij} = \cos\theta_{ij}$). The condition number $k(X^T X) \approx 9.56 \cdot 10^4$ and median *VIF* = 22.4 confirm severe multicollinearity in the HAR data.

Singular value decomposition $X = U \Sigma V^T$ is simultaneously the algebraic foundation of PCA ($C = V \Lambda V^T$, $\lambda_j = \sigma_j^2/(m-1)$) and a tool for optimal compression. Projection onto *k* = 20 components eliminates multicollinearity ($Z_k^T Z_k = \Lambda_k$), retains 98.6 % of variance, and improves accuracy from 62.8 % to 65.5 %. The effective rank $eff_{rank} \approx 16.5$ confirms that the HAR data are effectively low-dimensional despite the formal dimensionality of $p = 561$.

References

1. Bellman, R.E. Adaptive Control Processes: A Guided Tour. – Princeton University Press, 1961. – 255 p.

2. Anguita, D. [et al.] A Public Domain Dataset for Human Activity Recognition Using Smartphones // ESANN 2013. – 2013. – P. 437–442.

3. Pedregosa, F. [et al.] Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

4. Golub, G. H., Van Loan, C. F. Matrix Computations. — 4th ed. — Johns Hopkins University Press, 2013. – 780 p.

5. Eckart, C., Young, G. The Approximation of One Matrix by Another of Lower Rank // Psychometrika. – 1936. – Vol. 1, No. 3. – P. 211–218.