A Generalization of the Pearson Correlation to Riemannian Manifolds

P. Michl \( \text{orcid.org/0000-0002-6398-0654} \)

The increasing application of deep-learning is accompanied by a shift towards highly non-linear statistical models. In terms of their geometry it is natural to identify these models with Riemannian manifolds. The further analysis of the statistical models therefore raises the issue of a correlation measure, that in the cutting planes of the tangent spaces equals the respective Pearson correlation and extends to a correlation measure that is normalized with respect to the underlying manifold. In this purpose the article reconstitutes elementary properties of the Pearson correlation to successively derive a linear generalization to multiple dimensions and thereupon a nonlinear generalization to principal manifolds, given by the Riemann-Pearson Correlation.

Key words: Nonlinear Correlation, Riemann-Pearson Correlation, Principal Manifold

1 Introduction

A fundamental issue, that accompanies the analysis of multivariate data, concerns the quantification of statistically dependency structures by association measures. Many approaches in this direction can be traced back to the late 19th century, where the issue was closely related to the task, to extract laws of nature from two dimensional scatter plots. This in particular applies to the widespread Pearson correlation coefficient.

Definition (Pearson Correlation). Let \( X: \Omega \to \mathbb{R} \) and \( Y: \Omega \to \mathbb{R} \) be random variables with finite variances \( \sigma_X^2 \) and \( \sigma_Y^2 \). Then the Pearson correlation \( \rho_{X,Y} \) is defined by:

\[
\rho_{X,Y} := \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1.1}
\]

Due to its popularity and simplicity the Pearson correlation has been generalized to a variety of different domains of application, including generic monotonous relationships, relationships between sets of random variables and asymmetric relationships (Zheng et al. 2010). In the purpose to provide a generalization to smooth curves and submanifolds, that allow an incorporation of structural assumptions, some elementary considerations have to be taken into account, that allow a separation between the pairwise quantification of dependencies and their global modelling.

2 Correlation and Regression Dilution

Pearson’s original motivation, was the regression of a straight line, that minimizes the averaged Euclidean distance to points, that are scattered about it (Pearson 1901, p561). Thereby his investigations were preceded by the observation, that for a measurement series the assumed “direction of causality” influences the estimate of the slope of the regression line. Thereby the direction of causality is implicated by the choice of an error model, that assumes one random variable to be error free and the other to account for the whole observed error. Pearson empirically observed, that for \( n \in \mathbb{N} \) points, given by i.i.d. realizations \( x \in \mathbb{R}^n \) of \( X \) and \( y \in \mathbb{R}^n \) of \( Y \), the least squares regression line of \( y \) on \( x \) only equals the regression line of \( x \) on \( y \), if all points perfectly fit on a straight line. In all other cases, however, the slopes of the respective regression lines turned out, not to be reciprocal and their product was found within the interval \([0,1]\). This observation was decisive for Pearson’s definition of the correlation coefficient. Thereby \( \rho_{X,Y} \) is estimated by its empirical counterpart \( \hat{\rho}_{xy} \), that replaces variances by sample variances and the covariance by the sample variance.

Lemma 1. Let \( X: \Omega \to \mathbb{R} \) and \( Y: \Omega \to \mathbb{R} \) be random variables with \( n \in \mathbb{N} \) i.i.d. realizations \( x \in \mathbb{R}^n \) and \( y \in \mathbb{R}^n \). Furthermore let \( \beta_x, \beta_y \in \mathbb{R} \) denote the slope of the linear regression of \( y \) on \( x \) and \( \beta_y \in \mathbb{R} \) the slope of the linear regression of \( x \) on \( y \). Then:

\[
\rho^2_{xy} = \beta_x \beta_y \tag{2.1}
\]

Proof of Lemma 1. The following proof is based on (Kenny et al. 1962). The least squares regression of \( y \) on \( x \) implicates, that for regression coefficients \( \alpha_x, \beta_x, \beta_z \in \mathbb{R} \) and a normal distributed random error \( \varepsilon \approx Y - (\beta_x X + \alpha_x) \) the log-likelihood of the realizations is maximized, if and only if the \( \ell^2 \)-norm of the realizations of \( \varepsilon \) is minimized, such that:

\[
SSE_y(\alpha_x, \beta_x) := \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))^2 \to \min \tag{2.2}
\]

Since \( SSE_y \) is a quadratic function of \( \alpha_x \) and \( \beta_x \) and therefore convex, it has a unique global minimum at:

\[
\frac{\partial}{\partial \alpha_x} SSE_y = 2 \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))(-x_i) = 0 \tag{2.3}
\]

\[
\frac{\partial}{\partial \beta_x} SSE_y = 2 \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))(-1) = 0 \tag{2.4}
\]

\[
\sigma_X^2 \sigma_Y^2 \tag{2.1}
\]

\[
\rho_{X,Y} := \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1.1}
\]
By equating the coefficients, equations 2.3 and 2.4 can be rewritten as a system of linear equations of \(\alpha_x\) and \(\beta_x\):

\[
\alpha_x n + \beta_x \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i
\]

(2.5)

\[
\alpha_x \sum_{i=1}^{n} x_i + \beta_x \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i
\]

(2.6)

Consequently in matrix notation the vector \((\alpha_x, \beta_x)^T\) is determined by:

\[
\begin{pmatrix}
\alpha_x \\
\beta_x 
\end{pmatrix} = \left( \frac{n}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} x_i \right)^{-1} \left( \sum_{i=1}^{n} x_i y_i \right)
\]

(2.7)

Let \(\bar{x}, \bar{y}\) respectively denote the sample means. Then by calculating the matrix inverse, the slope \(\beta_x\) equates to:

\[
\beta_x = \left( \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \right) \left( \sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \right)^{-1}
\]

(2.8)

Thereupon by substituting the sample variance:

\[
\sigma_x^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

\[
= \frac{1}{n} \left( \sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \right)
\]

(2.9)

And the sample covariance:

\[
\text{Cov}(x, y) := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}
\]

(2.10)

It follows from equation 2.8, that:

\[
\beta_x = \frac{\text{Cov}(x, y)}{\sigma_x^2}
\]

(2.11)

Conversely the slope \(\beta_y\) of the linear regression of \(x\) on \(y\) mutatis mutandis equates to:

\[
\beta_y = \frac{\text{Cov}(y, x)}{\sigma_y^2}
\]

(2.12)

By the symmetry of Cov it the follows, from equations 2.11 and 2.12 that:

\[
\beta_x \beta_y = \frac{\text{Cov}(x, y)^2}{\sigma_x^2 \sigma_y^2} = \rho_{xy}^2
\]

(2.13)

Lemma 1 shows, that \(\rho_{x,y}\) may be regarded as the geometric mean of the regression slopes \(\beta_x\) and \(\beta_y\), where \(\beta_x\) and \(\beta_y\) respectively describe the causal relationships \(X \rightarrow Y\) and \(Y \rightarrow X\). Thereby \(X\) and \(Y\) respectively are treated as error free regressor variables to predict the corresponding response variable, that captures the overall error. The mutual linear relationship \(X \leftrightarrow Y\) is then described by a regression line, that equally treats errors in both variables. As an immediate consequence of this symmetry it follows, that this total least squares regression line is unique, and its slope \(\beta_x^*\), that describes \(y\) by \(x\) is reciprocal to the slope \(\beta_y^*\), that describes \(x\) by \(y\) such that \(\beta_x^* \beta_y^* = 1\). In this sense \(\beta_x\) and \(\beta_y\) may be regarded as biased estimations of \(\beta_x^*\) and \(\beta_y^*\). Thereby the bias generally is known as “regression dilution” or “regression attenuation”. For the case that both errors are independent and normal distributed, this bias can be corrected by a prefactor, that incorporates the error of the respective regressor variable. An application of this correction to lemma 1 then shows, that \(\rho_{x,y}\) has a consistent estimations by the sample variances of \(x\) and \(y\) and the variances of their respective errors \(\varepsilon_X\) and \(\varepsilon_Y\).

**Proposition 2.** Let \(X : \Omega \rightarrow \mathbb{R}\) and \(Y : \Omega \rightarrow \mathbb{R}\) be random variables with \(n \in \mathbb{N}\) i.i.d. realizations \(x \in \mathbb{R}^n\) and \(y \in \mathbb{R}^n\) and random errors \(\varepsilon_X \sim N(0, \eta_X^2)\) and \(\varepsilon_Y \sim N(0, \eta_Y^2)\). Then:

\[
\rho_{x,y}^2 \xrightarrow{p} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right) \left(1 - \frac{\eta_Y^2}{\sigma_y^2}\right), \text{ for } n \rightarrow \infty
\]

(2.14)

**Proof of Proposition 2.** Let \(\beta_x \in \mathbb{R}\) be the slope of the ordinary least squares (OLS) regression line of \(y\) on \(x\), where \(x\) is assumed to realize \(X\) with a normal distributed random error \(\varepsilon_X \sim N(0, \eta_X^2)\). Then \(X\) decomposes into (i) an unobserved error free regressor variable \(X^*\) and (ii) the random error \(\varepsilon_X\), such that:

\[
X \sim X^* + \varepsilon_X
\]

(2.15)

With respect to this decomposition, the slope \(\beta_x^*\) of the total least squares (TLS) regression line, that also considers \(\varepsilon_X\), then is identified by the slope of the OLS regression of \(y\) on \(x^*\), where \(x^*\) realizes \(X^*\). Thereupon let \(\sigma_x^2\) be the empirical variance of \(x^*\), then according to (Snedecor et al. 1967) it follows, that:

\[
\beta_x \xrightarrow{p} \frac{\sigma_x^2}{\sigma_x^2 + \eta_X^2} \beta_x^*, \text{ for } n \rightarrow \infty
\]

(2.16)

Since furthermore \(\varepsilon_X\) by definition is statistically independent from \(X^*\), it can be concluded, that:

\[
\sigma_x^2 = \text{Var}(X^* + \varepsilon_X) = \text{Var}(X^*) + \text{Var}(\varepsilon_X) = \sigma_x^2 + \eta_X^2
\]

(2.17)

Such that:

\[
\frac{\sigma_x^2}{\sigma_x^2 + \eta_X^2} \xrightarrow{2.17} \frac{\sigma_x^2 - \eta_X^2}{\sigma_x^2} = 1 - \frac{\eta_X^2}{\sigma_x^2}
\]

(2.18)

And therefore by equation 2.16 that:

\[
\beta_x \xrightarrow{p} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right) \beta_x^*, \text{ for } n \rightarrow \infty
\]

(2.19)

Conversely let now \(\beta_x^* \in \mathbb{R}\) be the the slope of the OLS regression of \(x\) on \(y\), where \(y\) is assumed to realize \(Y\) with a random error \(\varepsilon_Y \sim N(0, \eta_Y^2)\). Then also the corrected slope \(\beta_y^*\) mutatis mutandis satisfies the relation given by
equation 2.19 and by the representation of \( \rho_{x,y} \), as given by lemma 1, it then can be concluded, that:

\[
\rho^2_{x,y} \leq \beta_x \beta_y
\]  

(2.20)

The proposition then follows by the uniqueness of the total least squares regression line for known variances \( \eta^2_X \) and \( \eta^2_Y \), such that:

\[
\beta_y^* = \frac{1}{\beta_x^*}
\]

3 A Generalization to linear Principal Manifolds

Within the same publication, in which Pearson introduced the correlation coefficient, he also developed a structured approach that determines the straight line, that minimizes the Euclidean distance (Pearson 1901, p563). His method, which later received attribution as the method of Principal Component Analysis (PCA), however, even went further and allowed a canonical generalization of the problem in the following sense: For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be a multivariate random vector and for \( n \in \mathbb{N} \) let \( x \in \mathbb{R}^{n \times d} \) be an i.i.d. realization of \( X \). Then for any given \( k \in \mathbb{N} \) with \( k \leq d \) the goal is, to determine an affine linear subspace \( L \subseteq \mathbb{R}^d \) of dimension \( k \), that minimizes the summed Euclidean distance to \( x \). In order to solve this problem, the fundamental idea of Pearson was, to transfer the principal axis theorem from ellipsoids to multivariate Gaussian distributed random vectors. Thereupon, however, the method also can be formulated with respect to generic elliptical distributions.

**Definition (Elliptical Distribution).** For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be a random vector. Then \( X \) is elliptically distributed, if there exists a random vector \( S: \Omega \to \mathbb{R}^k \) with \( k \leq d \), which distribution is invariant to rotations, a matrix \( A \in \mathbb{R}^{d \times k} \) of rank \( k \) and a vector \( b \in \mathbb{R}^d \), such that:

\[
X \sim AS + b
\]  

(3.1)

Consequently a random vector \( X \) is elliptically distributed, if it can be represented by an affine transformation of a radial symmetric distributed random vector \( S \). The decisive property, that underpins the choice of elliptical distributions, lies within their coincidence of linear and statistical dependencies, which allows to decompose \( X \) in statistically independent components by a linear decomposition. This property allows, to substantiate the multidimensional “linear fitting problem” with respect to an orthogonal projection.

**Proposition 3.** For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be an elliptically distributed random vector, \( L \subseteq \mathbb{R}^d \) an affine linear subspace of \( \mathbb{R}^d \) and \( \pi_L \) the orthogonal projection of \( \mathbb{R}^d \) onto \( L \). Then the following statements are equivalent:

- (i) \( L \) minimizes the Euclidean distance to \( X \)
- (ii) \( \mathbb{E}(X) \in L \) and \( L \) maximizes the variance \( \text{Var}(\pi_L(X)) \)

**Proof.** Let \( Y_L := X - \pi_L(X) \), then the Euclidean distance between \( X \) and \( L \) can be written:

\[
d(X, \pi_L(X))^2 = \mathbb{E}(\|X - \pi_L(X)\|^2) = \mathbb{E}(Y_L^2)
\]  

(3.2)

This representation can furthermore be decomposed by using the algebraic formula for the variance:

\[
\mathbb{E}(Y_L^2) = \text{Var}(Y_L) + \mathbb{E}(Y_L)^2
\]  

(3.3)

Let now be \( Y_L^\perp := \pi_L(X) \), then \( X = Y_L + Y_L^\perp \) and \( Y_L \) and \( Y_L^\perp \) are uncorrelated, such that:

\[
\text{Var}(X) = \text{Var}(Y_L + Y_L^\perp) = \text{Var}(Y_L) + \text{Var}(Y_L^\perp)
\]  

(3.4)

From equations 3.2, 3.3 and 3.4 it follows that:

\[
d(X, \pi_L(X))^2 = \mathbb{E}(X) - \text{Var}(Y_L^\perp) + \mathbb{E}(Y_L)^2
\]  

(3.5)

Consequently the Euclidean distance is minimized, if and only if the right side of equation 3.5 is minimized. The first term \( \mathbb{E}(X) \), however, does not depend on \( L \) and since \( X \) is elliptically distributed, the linear independence of \( Y_L^\perp \) and \( Y_L \) is sufficient for statistically independence. It follows, that the Euclidean distance is minimized, if and only if: (1) The term \( \mathbb{E}(Y_L)^2 \) is minimized and (2) the term \( \text{Var}(Y_L^\perp) \) is maximized. Concerning (1) it follows, that:

\[
\mathbb{E}(Y_L)^2 = \mathbb{E}(X) - \pi_L(X))^2 = (\mathbb{E}(X) - \pi_L(\mathbb{E}(X)))^2
\]

Therefore the term \( \mathbb{E}(Y_L)^2 \) is minimized, if and only if \( \pi_L(\mathbb{E}(X)) = \mathbb{E}(X) \), which in turn means that \( \mathbb{E}(X) \in L \). Concerning (2), the proposition immediately follows by the definition of \( Y_L^\perp \).

Let now be \( k \leq d \). In order to derive an affine linear subspace \( L \subseteq \mathbb{R}^d \) that minimizes the Euclidean distance to \( X \), proposition ... states, that is suffices to provide an \( L \) which (1) is centred in \( X \), such that \( \mathbb{E}(X) \in L \), and (2) maximizes the variance of the projection. In order to maximize \( \text{Var}(\pi_L(X)) \), however, it is beneficial to give a further representation.

**Lemma 4.** For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be an elliptically distributed random vector, \( L \subseteq \mathbb{R}^d \) an affine linear subspace of \( \mathbb{R}^d \), which for an \( k \leq d \), a vector \( v \in \mathbb{R}^d \) and an orthonormal basis \( u_1, \ldots, u_k \in \mathbb{R}^d \) is given by:

\[
L = v + \bigoplus_{i=1}^k \mathbb{R}u_i
\]
Let further be \( \pi_L : \mathbb{R}^d \rightarrow L \) the orthogonal projection of \( \mathbb{R}^d \) onto \( L \). Then the variance of the projection is given by:

\[
\text{Var}(\pi_L(X)) = \sum_{i=1}^{k} u_i^T \text{Cov}(X) u_i
\]

Proof. Let \( L' := L + \mathbb{E}(X) - \mathbf{v} \), then the orthogonal projection \( \pi_{L'}(X) \) decomposes into individual orthogonal projections to the respective basis vectors, such that:

\[
\pi_{L'}(X) = \mathbb{E}(X) + \sum_{i=1}^{k} \langle X - \mathbb{E}(X), u_i \rangle u_i
\]

Let \( \tilde{X}_i := (X, u_i) u_i \), for \( i \in \{1, \ldots, k\} \). The total variance of this projection is then given by:

\[
\text{Var}(\pi_{L'}(X)) = \text{Var}\left( \mathbb{E}(X) + \sum_{i=1}^{k} \tilde{X}_i - \sum_{i=1}^{k} \mathbb{E}(\tilde{X}_i) \right)
\]

\[
= \text{Var}\left( \sum_{i=1}^{k} \tilde{X}_i \right)
\]

Since the random variables \( \tilde{X}_i \) by definition are uncorrelated, the algebraic formula for the variance can be used to decompose the variance:

\[
\text{Var}\left( \sum_{i=1}^{k} \tilde{X}_i \right) = \sum_{i=1}^{k} \text{Var}(\tilde{X}_i)
\]

By equating the term \( \text{Var}(\tilde{X}_i) \), for \( i \in \{1, \ldots, k\} \) it follows, that:

\[
\text{Var}(\tilde{X}_i) = \text{Var}\left((X, u_i) u_i\right) = \text{Var}\left(X^T u_i\right) u_i^2 = \text{Var}\left(X^T u_i\right)
\]

And furthermore by introducing the covariance matrix \( \text{Cov}(X) \):

\[
\text{Var}\left(X^T u_i\right) := \mathbb{E}\left((X^T u_i)^T (X^T u_i)\right) = u_i^T \mathbb{E}(X^T X) u_i \equiv u_i^T \text{Cov}(X) u_i
\]

Summarized the equations 3.7, 3.8, 3.9 and 3.10 provide a representation for the variance of the projection to \( L' \):

\[
\text{Var}(\pi_{L'}(X)) = \sum_{i=1}^{k} u_i^T \text{Cov}(X) u_i
\]

Finally the total variance of the projection is invariant under translations of \( L \), such that:

\[
\text{Var}(\pi_L(X)) = \text{Var}(\pi_{L'}(X) + \mathbb{E}(X) - \mathbf{v}) = \text{Var}(\pi_{L'}(X))
\]

\[
= 3.8 \sum_{i=1}^{k} u_i^T \text{Cov}(X) u_i
\]

Lemma 4, shows, that for elliptically distributed random vectors \( X \) the best fitting linear subspaces are completely determined by the expectation \( \mathbb{E}(X) \) and the covariance matrix \( \text{Cov}(X) \). On this point it is important to notice, that the covariance matrix is symmetric, which allows its diagonalization with regard to real valued Eigenvalues.

**Lemma 5.** For \( d \in \mathbb{N} \) let \( X : \Omega \rightarrow \mathbb{R}^d \) be an elliptically distributed random vector, \( L \subseteq \mathbb{R}^d \) an affine linear subspace of \( \mathbb{R}^d \), which for an \( k \leq d \), a vector \( \mathbf{v} \in \mathbb{R}^d \) and an orthonormal basis \( u_1, \ldots, u_k \in \mathbb{R}^d \) is given by:

\[
L = \mathbf{v} + \bigoplus_{i=1}^{k} \mathbb{R} u_i
\]

Let further be \( \pi_L : \mathbb{R}^d \rightarrow L \) the orthogonal projection of \( \mathbb{R}^d \) onto \( L \), as well as \( \lambda_1, \ldots, \lambda_d \in \mathbb{R} \) the eigenvalues of \( \text{Cov}(X) \). Then there exist numbers \( a_1, \ldots, a_d \in [0, 1] \) with \( \sum_{i=1}^{d} a_i = k \), such that:

\[
\text{Var}(\pi_L(X)) = \sum_{i=1}^{d} \lambda_i a_i
\]

Proof. From lemma 4 it follows, that:

\[
\text{Var}(\pi_L(X)) = \sum_{j=1}^{k} u_j^T \text{Cov}(X) u_j
\]

Since the covariance matrix \( \text{Cov}(X) \) is a symmetric matrix, there exists an orthonormal basis transformation matrix \( S \in \mathbb{R}^{d \times d} \) and a diagonal matrix \( D \in \mathbb{R}^{d \times d} \), such that \( \text{Cov}(X) = S^T D S \). Then the variance \( \text{Var}(\pi_L(X)) \) has a decomposition, given by:

\[
\text{Var}(\pi_L(X)) = \sum_{i=1}^{k} u_j^T S^T D S u_j
\]

\[
= \sum_{i=1}^{k} (S u_j)^T D S u_j
\]

For \( j \in \{1, \ldots, k\} \) let now \( c_j := S u_j \) and for \( i \in \{1, \ldots, n\} \) let the number \( a_i \in \mathbb{R} \) be defined by:

\[
a_i := \sum_{j=1}^{k} (c_{ji})^2
\]

Then according to Lemma 4 the variance \( \text{Var}(\pi_L(X)) \) can be decomposed:

\[
\sum_{i=1}^{k} c_{ji} D c_j = \sum_{j=1}^{k} \sum_{i=1}^{d} c_{ji} \lambda_i e_{ji}
\]

\[
= \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{k} (c_{ji})^2
\]

\[
= \sum_{i=1}^{d} \lambda_i a_i
\]
Furthermore since \( u_1, \ldots, u_k \) is an orthonormal basis and \( S \) an orthonormal matrix it follows that also \( c_1, \ldots, c_k \) is an orthonormal basis. Consequently for \( i \in \{1, \ldots, d\} \) it holds, that:

\[
a_i = \sum_{j=1}^k (c_{ji})^2 \leq \sum_{j=1}^k \| c_j \|_2 \leq 1
\]

And furthermore by its definition it follows, that \( a_i \geq 0 \), such that \( a_i \in [0, 1] \). Besides this the sum over all \( a_i \) equates to:

\[
\sum_{i=1}^d a_i = \sum_{i=1}^d \sum_{j=1}^k (c_{ji})^2 = \sum_{j=1}^k c_j^T c_j = k
\]

Accordingly the choice \( u_j = c_j \) for \( i \in \{1, \ldots, k\} \) maximizes \( \text{Var}(\pi_L(X)) \) and \( L \) has a representation, given by:

\[
L = \mathbb{E}(X) + \bigoplus_{i=1}^k \mathbb{R}c_i
\]

Then according to Lemma 5 there exist numbers \( a_1, \ldots, a_d \in [0, 1] \) with \( \sum_{i=1}^d a_i = k \), such that:

\[
\text{Var}(\pi_L(X)) = \sum_{i=1}^d \lambda_i a_i
\]

Thereupon (ii) is satisfied, if and only if the numbers \( a_i \) maximize this sum. Since the covariance matrix \( \text{Cov}(X) \) is positive semi-definite, the eigenvalues \( \lambda_i \) are not negative such that the sum is maximized for:

\[
a_i = \begin{cases} 
1 & \text{for } i \in \{1, \ldots, k\} \\
0 & \text{else}
\end{cases}
\]

Such that:

\[
\sum_{j=1}^k u_j^T \text{Cov}(X) u_j = \sum_{i=1}^d \lambda_i a_i = \sum_{i=1}^k \lambda_i = \sum_{j=1}^k c_j^T \text{Cov}(X) c_j
\]

Definition (Linear Principal Manifold). For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be a random vector. Then a vector \( c \in \mathbb{R}^d \) with \( c \neq 0 \) is a principal component for \( X \), iff there exists an \( \lambda \in \mathbb{R} \), such that:

\[
\text{Cov}(X) \cdot c = \lambda c
\]

Furthermore let \( L \subset \mathbb{R}^d \) be an affine linear subspace of \( \mathbb{R}^d \) with dimension \( k \leq d \). Then \( L \) is a linear \( k \)-principal manifold for \( X \), if there exists a set set \( c_1, \ldots, c_k \) of linear independent principal components for \( X \), such that:

\[
L = \mathbb{E}(X) + \bigoplus_{i=1}^k \mathbb{R}c_i
\]

Then \( L \) is termed maximal, iff the sum of the Eigenvalues \( \lambda_1, \ldots, \lambda_k \), that correspond to the principal components \( c_1, \ldots, c_k \) is maximal.

Proposition 6. For \( d \in \mathbb{N} \) let \( X: \Omega \to \mathbb{R}^d \) be an elliptically distributed random vector and \( L \subset \mathbb{R}^d \) an affine linear subspace. Then the following statements are equivalent:

(i) \( L \) minimizes the Euclidean distance to \( X \)
(ii) \( L \) is a maximal linear principal manifold for \( X \)

Proof. “\( \Rightarrow \)” Let \( \pi_L: \mathbb{R}^d \mapsto L \) denote the orthogonal projection of \( \mathbb{R}^d \) onto \( L \). Then according to proposition 3 \( L \) minimizes the averaged Euclidean distance to \( X \), if and only if (i) \( \mathbb{E}(X) \in L \) and (ii) \( L \) maximizes the variance \( \text{Var}(\pi_L(X)) \). In particular (i) is satisfied, if and only if an orthonormal basis \( u_1, \ldots, u_k \in \mathbb{R}^d \) can be chosen, such that:

\[
L = \mathbb{E}(X) + \bigoplus_{i=1}^k \mathbb{R}u_i
\]

“\( \Leftarrow \)” Let \( \pi_L: \mathbb{R}^d \mapsto L \) denote the orthogonal projection of \( \mathbb{R}^d \) onto \( L \). Then according to proposition 3 \( L \) minimizes the averaged Euclidean distance to \( X \), if and only if (i) \( \mathbb{E}(X) \in L \) and (ii) \( L \) maximizes the variance \( \text{Var}(\pi_L(X)) \). In particular (i) is satisfied, if and only if an orthonormal basis \( u_1, \ldots, u_k \in \mathbb{R}^d \) can be chosen, such that:

\[
L = \mathbb{E}(X) + \bigoplus_{i=1}^k \mathbb{R}u_i
\]
Such that by the definition of the reliability it follows, that:

\[ R_i = 1 - \frac{\hat{\sigma}^2_{x_i}}{\sigma^2_{x_i}} \]

Consequently:

\[ \rho^2_{x_i,x_j|L} = \left(1 - \frac{\hat{\sigma}^2_{x_i}}{\sigma^2_{x_i}}\right) \left(1 - \frac{\hat{\sigma}^2_{x_j}}{\sigma^2_{x_j}}\right) \]

With \( n \in \mathbb{N} \) i.i.d. realizations \( x \in \mathbb{R}^{n \times 2} \) of \( X \) an empirical \( L \)-Correlation \( \rho^2_{x_i,x_j|L} \) is then given by replacing the variances by the sample variances. Then by proposition 2 it follows, that:

\[ \rho^2_{x_i,x_j} \xrightarrow{P} \rho^2_{x_i,x_j|L}, \text{ for } n \to \infty \]

\[ \square \]

4 The Riemann-Pearson Correlation

Linear principal manifolds allow the projection of a random vector \( X : \Omega \to \mathbb{R}^d \) onto a linear subspace \( L \subseteq \mathbb{R}^d \), which maximally preserves the linear dependency structure of \( X \) in terms of its covariances. Thereby for the orthogonal projection \( \pi_L : \mathbb{R}^d \to L \), the variance on \( L \), given by \( \text{Var}(\pi_L(X)) \), is referred as the explained variance and the orthogonal deviation \( \text{Var}(X - \pi_L(X)) \) as the unexplained variance. Thereupon by the assumption, that \( X \) is elliptically distributed, it can be concluded, that linear independence coincides with statistically independence, that is \( \pi_L(X) \) and \( X - \pi_L(X) \) are statistically independent and therefore allow the following decomposition:

\[ \text{Var}(X) = \text{Var}(\pi_L(X)) + \text{Var}(X - \pi_L(X)) \]

This decomposition, as shown by theorem 7, is of fundamental importance for the correlation over linear Principal Manifolds, since it determines the reliabilities of the respective random variable \( X \) by the ratio:

\[ R = 1 - \frac{\text{explained variance}}{\text{total variance}} \]

On this point of the discussion it’s just a small step to generalize the principal components, by a smooth curves \( \gamma : [a, b] \to \mathbb{R}^d \) (figure 4.1). This is particular appropriate, if the assumption of an elliptically distribution can only hardly be justified, like for observed dynamical systems. Thereby the evolution function generates a smooth submanifold \( M \subseteq \mathbb{R}^d \) within the observation space \( \mathbb{R}^d \), and an “error free” observation can be identified by a random vector \( X^* \), with outcomes on \( M \). Additionally, however, the observation function may be regarded to be subjected to a measurement error \( \varepsilon \). By the assumption, that \( \varepsilon \) has an elliptical distribution, then the distribution of the observable random vector \( X \) is represented by a elliptical \( M \)-distribution.

**Definition (\( M \)-Distribution).** For \( d \in \mathbb{N} \) let \( X^* : \Omega \to \mathbb{R}^d \) be a random vector and \( M \subseteq \mathbb{R}^d \) a smooth \( k \)-submanifold of \( \mathbb{R}^d \) with \( k \leq d \). Then \( X^* \) is \( M \)-distributed, iff for the probability density \( P \), which is induced by \( X^* \), it holds, that:

\[ P(X^* = x) > 0 \Leftrightarrow x \in M \quad (4.1) \]

Thereupon a random vector \( X : \Omega \to \mathbb{R}^d \) is elliptically \( M \)-distributed, iff there exists an \( M \)-distributed random vector \( X^* : \Omega \to \mathbb{R}^d \) and an elliptically distributed random error \( \varepsilon : \Omega \to \mathbb{R}^d \), such that:

\[ X \sim X^* + \varepsilon \quad (4.2) \]

The assumption, that the observed random vector \( X \), is elliptically \( M \)-distributed, is very general, but allows an estimation of \( M \) by minimizing the averaged Euclidean distance to \( X \). Thereby the tangent spaces \( T_xM \) have a basis, given by \( k \) principal components of local infinitesimal covariances, such that the remaining \( d - k \) principal components describe the normal space \( N_xM \), which is orthogonal to the tangent space \( T_xM \). Since \( T_xM \) and \( N_xM \) are equipped with an induced Riemannian metric, which is simply given by the standard scalar product, there exists a minimal orthogonal projection \( \pi_M : \mathbb{R}^d \to M \), that maps any realization \( x \) of \( X \) to a
closest point on $\mathcal{M}$. Then proposition 3 motivates properties for $\mathcal{M}$ to minimize the averaged Euclidean distance to realizations of $\mathbf{X}$. This provides the definition of smooth $k$-principal manifolds (Hastie et al. 1989, p513).

**Definition (Principal Manifold).** For $d \in \mathbb{N}$ let $\mathbf{X}: \Omega \to \mathbb{R}^d$ be a random vector, $\mathcal{M} \subseteq \mathbb{R}^d$ a (smooth) $k$-submanifold of $\mathbb{R}^d$ with $k \leq d$ and $\pi_{\mathcal{M}}: \mathbb{R}^d \hookrightarrow \mathcal{M}$ a minimal orthogonal projection onto $\mathcal{M}$. Then $\mathcal{M}$ is a (smooth) $k$-principal manifold for $\mathbf{X}$, iff $\forall x \in \mathcal{M}$ it holds, that:

$$
\mathbb{E}(\mathbf{X} \in \pi_{\mathcal{M}}^{-1}(x)) = x
$$

Furthermore $\mathcal{M}$ is termed maximal, iff $\mathcal{M}$ maximizes the explained variance $\text{Var}(\pi_{\mathcal{M}}(\mathbf{X}))$.

By extending the local properties of the tangent spaces to the underlying manifold, by propositions 3 and 6 it can be concluded, that maximal principal manifolds minimize the Euclidean distance to $\mathbf{X}$. Intuitively this can be understood as follows: The principal manifold property assures, that:

$$
\text{Var}(\mathbf{X}) = \text{Var}(\pi_{\mathcal{M}}(\mathbf{X})) + \text{Var}(\mathbf{X} - \pi_{\mathcal{M}}(\mathbf{X}))
$$

Consequently the choice of $\mathcal{M}$ maximizes $\text{Var}(\pi_{\mathcal{M}}(\mathbf{X}))$ if and only if it minimizes $\text{Var}(\mathbf{X} - \pi_{\mathcal{M}}(\mathbf{X}))$, which equals the variance of the error and therefore the Euclidean distance. At closer inspection, however, it turns out, that in difference to linear principal manifolds, the maximization problem is ill-defined for arbitrary smooth principal manifolds, since for any finite number of realizations trivial solutions can be found by smooth principal manifolds, that interpolate the realizations and therefore provide a perfect explanation. In order to close this gap, further structural assumptions have to be incorporated, ether by a parametric family $\{f_g\}_{g \in G}$ that restricts the possible solutions - or by a regularization, as given in the elastic map algorithm that penalizes long distances and strong curvature (Gorban et al. 2008). Due to the complexity of this topic, however, it is left to the second chapter, where Energy based models are used to overcome this deficiency. In the following the generalization of the correlation to smooth principal manifolds for convenience is defined with respect to a principal manifold $\mathcal{M}$, which is maximal “with respect to appropriate restrictions”.

**Definition (Riemann-Pearson Correlation).** For $d \in \mathbb{N}$ let $\mathbf{X}: \Omega \to \mathbb{R}^d$ be a random vector, $\mathcal{M}$ a smooth principal manifold for $\mathbf{X}$, which is maximal “with respect to appropriate restrictions” and $\pi_{\mathcal{M}}: \mathbb{R}^d \hookrightarrow \mathcal{M}$ a minimal orthogonal projection. Then for any $i, j \in \{1, \ldots, d\}$ the Riemann-Pearson Correlation between $X_i$ and $X_j$ is given by:

$$
\rho_{X_i, X_j | \mathcal{M}}^2 := R_i R_j \int_{\mathcal{M}} S_{i,j}(x) S_{j,i}(x) \, dP_{\mathcal{M}}
$$

where for $i \in \{1, \ldots, d\}$ the reliability of $X_i$ with respect to $\mathcal{M}$ is given by:

$$
R_i := 1 - \frac{\text{Var}_i(\mathbf{X} - \pi_{\mathcal{M}}(\mathbf{X}))}{\text{Var}_i(\mathbf{X})}
$$

and for $i, j \in \{1, \ldots, d\}$ the local sensitivity of $X_i$ with respect to $X_j$ by:

$$
S_{i,j}(x) := \left. \frac{\partial}{\partial x_j} (x - \pi_{\mathcal{M}}(x)) \right|_{x} \quad (4.6)
$$

**Proposition 8.** For an elliptical $\mathcal{M}$–distributed random vector $\mathbf{X}$ the Riemann-Pearson Correlation generalizes the $L$-Correlation to smooth principal manifolds.

Proof. Let $L$ be a maximal linear principal manifold for $\mathbf{X}: \Omega \to \mathbb{R}^d$, and $x$ a realization of $\mathbf{X}$ then there exists an $\beta \in \mathbb{R}$ with:

$$
S_{i,j}(x) = \frac{\partial}{\partial x_j} (x - \pi_L(x)) \quad (4.7)
$$

Furthermore for $\epsilon \neq 0$:

$$
S_{i,j}(x) = \frac{\partial}{\partial x_i} (x - \pi_L(x)) \quad (4.8)
$$

Such that $S_{i,j}(x)S_{j,i}(x) = 1$. Consequently for $\mathcal{M} = L$ it follows, that:

$$
\rho_{X_i, X_j | \mathcal{M}}^2 = R_i R_j \int_{\mathcal{M}} S_{i,j}(x) S_{j,i}(x) \, dP_{\mathcal{M}} = R_i R_j \int_{\mathcal{M}} dP_{\mathcal{M}} = R_i R_j = \rho_{X_i, X_j | \mathcal{M}}^2
$$



References

[1] Shurong Zheng, Ning-zhong Shi, and Zhengjun Zhang. Generalized Measures of Correlation. Manuscript, pages 1–45, 2010.

[2] Karl Pearson. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(1):559–572, 1901.

[3] J. Kenney and E. Keeping. Linear regression and correlation. Mathematics of statistics, 1:252–285, 1962.

[4] G. W. Snedecor and W. G. Cochran. Statistical Methods. Iowa State University Press, Ames, 6 edition, 1967.

[5] Trevor Hastie and Werner Stuetzle. Principal Curves. Journal of the American Statistical Association, 84(406):502–516, 1989.

[6] Alexander N. Gorban and Andrei Y. Zinovyev. Principal Graphs and Manifolds. Arxiv preprint arXiv08090490, page 36, 2008.