THE RIEMANN-PEARSON CORRELATION

P. MICHL <PATRICK.MICHL@GMAIL.COM>

Abstract. This Article generalizes the Pearson correlation to Riemannian manifolds. In this purpose the first section reconstitutes elementary properties of the Pearson correlation to derive a representation with respect to the a regression line. The second section introduces principal components to derive this respective line, and thereupon to generalize it to linear subspaces of it's embedding space. Thereby the theory is derived for generic elliptical distributions and the principal components are introduced as an orthogonal basis of the embedding space, that decorrelates elliptically distributed random vectors. In the subsequent section the spaces, that are spanned by principal components, are used to identify the tangent spaces of principal manifolds. As principal manifolds, however, are not assured to exist for arbitrary underlying densities, a class of smooth manifold based densities is introduced, that closes this gap. Finally with the Riemann-Pearson correlation is defined, which is shown to generalize the Pearson correlation to densities of this respective class

1. Inrtroduction

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 19^{th} 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 $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables with finite variances σ_X^2 and σ_Y^2. Then the Pearson correlation $\rho_{X,Y}$ is defined by:*

$$\rho_{X,Y} \coloneqq \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1.1}$$

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 $\boldsymbol{x} \in \mathbb{R}^n$ of X and $\boldsymbol{y} \in \mathbb{R}^n$ of Y, the **least squares** regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ only equals the regression line of $\boldsymbol{x}$ on $\boldsymbol{y}$, 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 $\rho_{x,y}$, that replaces variances by sample variances and the covariance by the sample variance.

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

$$\rho_{x,y}^2 = \beta_x \beta_y \tag{2.1}$$

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

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

Since SSE_y is a quadratic function of α_x and β_x and therefore convex, it has a unique global minimum at:

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

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

By equating the coefficients, equations 2.3 and 2.4 can be rewritten as a system of linear equations of α_x and β_x:

$$\alpha_x n + \beta_x \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.5}$$

$$\alpha_x \sum_{i=1}^{n} x_i + \beta_x \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{2.6}$$

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

$$\begin{pmatrix} \alpha_x \\ \beta_x \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{2.7}$$

Let $\overline{x}$, $\overline{y}$ respectively denote the *sample means*. Then by calculating the matrix inverse, the slope β_x equates to:

$$\beta_x = \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy} \right) \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right)^{-1} \tag{2.8}$$

Thereupon by substituting the *sample variance*:

$$\begin{aligned} \sigma_x^2 &\coloneqq \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) \end{aligned} \tag{2.9}$$

And the *sample covariance*:

$$\begin{aligned} \operatorname{Cov}(\boldsymbol{x},\, \boldsymbol{y}) &\coloneqq \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{1}{n} \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy} \right) \end{aligned} \tag{2.10}$$

It follows from equation 2.8, that:

$$\beta_x = \frac{\operatorname{Cov}(\boldsymbol{x},\, \boldsymbol{y})}{\sigma_x^2} \tag{2.11}$$

Conversely the slope β_y of the linear regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ mutatis mutandis equates to:

$$\beta_y = \frac{\operatorname{Cov}(\boldsymbol{y},\, \boldsymbol{x})}{\sigma_y^2} \tag{2.12}$$

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

$$\beta_x \beta_y = \frac{\operatorname{Cov}(\boldsymbol{x},\, \boldsymbol{y})^2}{\sigma_x^2 \sigma_y^2} = \rho_{xy}^2 \tag{2.13}$$

□

Lemma 1 shows, that $\rho_{x,y}$ may be regarded as the geometric mean of the regression slopes β_x and β_y, where β_x and β_y respectively describe the causal relationships $X \to Y$ and $Y \to X$. 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 $\beta_x^\star$, that describes $\boldsymbol{y}$ by $\boldsymbol{x}$ is reciprocal to the slope $\beta_y^\star$, that describes $\boldsymbol{x}$ by $\boldsymbol{y}$ such that $\beta_x^\star \beta_y^\star = 1$. In this sense β_x and β_y may be regarded as biased estimations of $\beta_x^\star$ and $\beta_y^\star$. 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ and the variances of their respective errors ε_X and ε_Y.

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

$$\rho_{x,y}^2 \xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right)\left(1 - \frac{\eta_Y^2}{\sigma_y^2}\right), \text{ for } n \to \infty \tag{2.14}$$

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

$$X \sim X^\star + \varepsilon_X \tag{2.15}$$

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

$$\beta_x \xrightarrow{P} \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \eta_X^2}\beta_x^\star, \text{ for } n \to \infty \tag{2.16}$$

Since furthermore ε_X by definition is statistically independent from $X^\star$, it can be concluded, that:

$$\begin{aligned} \sigma_x^2 &= \mathrm{Var}(X^\star + \varepsilon_X) \\ &= \mathrm{Var}(X^\star) + \mathrm{Var}(\varepsilon_X) = \sigma_{x^*}^2 + \eta_X^2 \end{aligned} \tag{2.17}$$

Such that:

$$\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \eta_X^2} \overset{2.17}{=} \frac{\sigma_x^2 - \eta_X^2}{\sigma_x^2} = 1 - \frac{\eta_X^2}{\sigma_x^2} \tag{2.18}$$

And therefore by equation 2.16 that:

$$\beta_x \xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right)\beta_x^\star, \text{ for } n \to \infty \tag{2.19}$$

Conversely let now $\beta_y \in \mathbb{R}$ be the the slope of the OLS regression of $\boldsymbol{x}$ on $\boldsymbol{y}$, where $\boldsymbol{y}$ is assumed to realize Y with a random error $\varepsilon_X \sim \mathcal{N}(0,\, \eta_Y^2)$. Then also the corrected slope $\beta_y^\star$ 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:

$$\begin{aligned} \rho_{x,y}^2 &\overset{1}{=} \beta_x \beta_x \\ &\xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right)\left(1 - \frac{\eta_Y^2}{\sigma_y^2}\right)\beta_x^\star \beta_y^\star, \text{ for } n \to \infty \end{aligned} \tag{2.20}$$

The proposition then follows by the uniqueness of the total least squares regression line for known variances η_X^2 and η_Y^2, such that:

$$\beta_y^\star = \frac{1}{\beta_x^\star}$$

□

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 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a multivariate random vector and for $n \in \mathbb{N}$ let $\boldsymbol{x} \in \mathbb{R}^{n\times d}$ be an i.i.d. realization of $\boldsymbol{X}$. 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 $\boldsymbol{x}$. 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 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a random vector. Then $\boldsymbol{X}$ is elliptically distributed, iff there exists a random vector $\boldsymbol{S}\colon \Omega \to \mathbb{R}^k$ 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 $\boldsymbol{b} \in \mathbb{R}^d$, such that:*

$$\boldsymbol{X} \sim A\boldsymbol{S} + \boldsymbol{b} \tag{3.1}$$

Consequently a random vector $\boldsymbol{X}$ is elliptically distributed, if it can be represented by an affine transformation of a radial symmetric distributed random vector $\boldsymbol{S}$. The decisive property, that underpins the choice of elliptical distributions, lies within their coincidence of linear and statistical dependencies, which allows to decompose $\boldsymbol{X}$ 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 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be an elliptically distributed random vector, $L \subseteq \mathbb{R}^d$ an affine linear subspace of $\mathbb{R}^d$ and π_L the orthogonal projection of $\mathbb{R}^d$ onto L. Then the following statements are equivalent:*

(i) *L minimizes the Euclidean distance to $\boldsymbol{X}$*

(ii) *$\mathbb{E}(\boldsymbol{X}) \in L$ and L maximizes the variance* $\mathrm{Var}(\pi_L(\boldsymbol{X}))$

Proof. Let $\boldsymbol{Y}_L \coloneqq \boldsymbol{X} - \pi_L(\boldsymbol{X})$, then the Euclidean distance between $\boldsymbol{X}$ and L can be written:

$$\begin{aligned} d(\boldsymbol{X}, \pi_L(\boldsymbol{X}))^2 &= \mathbb{E}(\|(\boldsymbol{X} - \pi_L(\boldsymbol{X}))\|_2^2) \\ &= \mathbb{E}(\boldsymbol{Y}_L^2) \end{aligned} \tag{3.2}$$

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

$$\mathbb{E}(\boldsymbol{Y}_L^2) = \mathrm{Var}(\boldsymbol{Y}_L) + \mathbb{E}(\boldsymbol{Y}_L)^2 \tag{3.3}$$

Let now be $\boldsymbol{Y}_L^\perp \coloneqq \pi_L(\boldsymbol{X})$, then $\boldsymbol{X} = \boldsymbol{Y}_L + \boldsymbol{Y}_L^\perp$ and $\boldsymbol{Y}_L$ and $\boldsymbol{Y}_L^\perp$ are uncorrelated, such that:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{X}) &= \mathrm{Var}(\boldsymbol{Y}_L + \boldsymbol{Y}_L^\perp) \\ &= \mathrm{Var}(\boldsymbol{Y}_L) + \mathrm{Var}(\boldsymbol{Y}_L^\perp) \end{aligned} \tag{3.4}$$

From equations 3.2, 3.3 and 3.4 it follows, that:

$$d(\boldsymbol{X}, \pi_L(\boldsymbol{X}))^2 = \mathrm{Var}(\boldsymbol{X}) - \mathrm{Var}(\boldsymbol{Y}_L^\perp) + \mathbb{E}(\boldsymbol{Y}_L)^2 \tag{3.5}$$

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

$$\begin{aligned}\mathbb{E}(\boldsymbol{Y}_L)^2 &= \mathbb{E}(\boldsymbol{X} - \pi_L(\boldsymbol{X}))^2 \\ &= (\mathbb{E}(\boldsymbol{X}) - \pi_L(\mathbb{E}(\boldsymbol{X})))^2\end{aligned}$$

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

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 $\boldsymbol{X}$, proposition ... states, that is suffices to provide an L which (1) is centred in $\boldsymbol{X}$, such that $\mathbb{E}(\boldsymbol{X}) \in L$, and (2) maximizes the variance of the projection. In order to maximize $\mathrm{Var}(\pi_L(\boldsymbol{X}))$, however, it is beneficial to give a further representation.

Lemma 4. *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ 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 $\boldsymbol{v} \in \mathbb{R}^d$ and an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ is given by:*

$$L = \boldsymbol{v} + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

Let further be $\pi_L \colon \mathbb{R}^d \to L$ the orthogonal projection of $\mathbb{R}^d$ onto L. Then the variance of the projection is given by:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i$$

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

$$\pi_{L'}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X}) + \sum_{i=1}^{k} \langle \boldsymbol{X} - \mathbb{E}(\boldsymbol{X}),\, \boldsymbol{u}_i \rangle\, \boldsymbol{u}_i \tag{3.6}$$

Let $\hat{X}_i \coloneqq \langle \boldsymbol{X},\, \boldsymbol{u}_i \rangle\, \boldsymbol{u}_i$, for $i \in \{1, \ldots, k\}$. The total variance of this projection is then given by:

$$\begin{aligned}\mathrm{Var}(\pi_{L'}(\boldsymbol{X})) &\overset{3.6}{=} \mathrm{Var}\left(\mathbb{E}(\boldsymbol{X}) + \sum_{i=1}^{k} \hat{X}_i - \sum_{i=1}^{k} \mathbb{E}(\hat{X}_i)\right) \\ &= \mathrm{Var}\left(\sum_{i=1}^{k} \hat{X}_i\right)\end{aligned} \tag{3.7}$$

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

$$\mathrm{Var}\left(\sum_{i=1}^{k} \hat{X}_i\right) = \sum_{i=1}^{k} \mathrm{Var}(\hat{X}_i) \tag{3.8}$$

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

$$\begin{aligned}\mathrm{Var}(\hat{X}_i) &= \mathrm{Var}\left(\langle \boldsymbol{X},\, \boldsymbol{u}_i \rangle\, \boldsymbol{u}_i\right) \\ &= \mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i\right) \boldsymbol{u}_i^2 \\ &= \mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i\right)\end{aligned} \tag{3.9}$$

And furthermore by introducing the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$:

$$\begin{aligned}\mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i\right) &\overset{\text{def}}{=} \mathbb{E}\left((\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i)^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i)\right) \\ &= \boldsymbol{u}_i^{\mathrm{T}} \mathbb{E}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}\right) \boldsymbol{u}_i \\ &\overset{\text{def}}{=} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i\end{aligned} \tag{3.10}$$

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

$$\mathrm{Var}(\pi_{L'}(\boldsymbol{X})) = \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i$$

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

$$\begin{aligned}\text{(3.11)}\quad \mathrm{Var}(\pi_L(\boldsymbol{X})) &= \mathrm{Var}(\pi_L(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{X}) - \boldsymbol{v}) \\ &= \mathrm{Var}(\pi_{L'}(\boldsymbol{X})) \\ &\overset{3.8}{=} \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i\end{aligned}$$

□

Lemma 4, shows, that for elliptically distributed random vectors $\mathbf{X}$ the best fitting linear subspaces are completely determined by the expectation $\mathbb{E}(\boldsymbol{X})$ and the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$. 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 $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ 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 $\boldsymbol{v} \in \mathbb{R}^d$ and an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ is given by:*

$$L = \boldsymbol{v} + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

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

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{d} \lambda_i a_i$$

Proof. From lemma 4 it follows, that:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{j=1}^{k} \boldsymbol{u}_j^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_j$$

Since the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ 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 $\mathrm{Cov}(\boldsymbol{X}) = S^T D S$. Then the variance $\mathrm{Var}(\pi_L(\boldsymbol{X}))$ has a decomposition, given by:

$$\begin{aligned}\mathrm{Var}(\pi_L(\boldsymbol{X})) &= \sum_{j=1}^{k} \boldsymbol{u}_j{}^{\mathrm{T}} S^{\mathrm{T}} D S \boldsymbol{u}_j \\ &= \sum_{i=1}^{k} (S\boldsymbol{u}_j)^{\mathrm{T}} D S \boldsymbol{u}_j\end{aligned}$$

For $j \in \{1, \ldots, k\}$ let now $\boldsymbol{c}_j \coloneqq S\boldsymbol{u}_j$ and for $i \in \{1, \ldots, n\}$ let the number $a_i \in \mathbb{R}$ be defined by:

$$a_i \coloneqq \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2$$

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

$$\begin{aligned}\sum_{i=1}^{k} \boldsymbol{c}_j{}^T D \boldsymbol{c}_j &= \sum_{j=1}^{k} \sum_{i=1}^{d} \boldsymbol{c}_{ji} \lambda_i \boldsymbol{c}_{ji} \\ &= \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 \\ &= \sum_{i=1}^{d} \lambda_i a_i\end{aligned}$$

Furthermore since $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ is an orthonormal basis and S an orthonormal matrix it follows that also $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ is an orthonormal basis. Consequentially for $i \in \{1, \ldots, d\}$ it holds, that:

$$a_i = \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 \leq \sum_{j=1}^{k} \|\boldsymbol{c}_{ji}\|_2 \leq 1$$

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:

$$\begin{aligned}\sum_{i=1}^{d} a_i &= \sum_{i=1}^{d} \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 \\ &= \sum_{j=1}^{k} \boldsymbol{c}_j^T \boldsymbol{c}_j = k\end{aligned}$$

□

With reference to the principal axis transformation, the eigenvectors of the covariance matrix are

then termed **principal components** and affine linear subspaces of the embedding space as **linear principal manifolds**.

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

$$\operatorname{Cov}(\boldsymbol{X}) \cdot \boldsymbol{c} = \lambda \boldsymbol{c} \tag{3.12}$$

Furthermore let $L \subseteq \mathbb{R}^d$ be an affine linear subspace of $\mathbb{R}^d$ with dimension $k \leq d$. Then L is a linear k-principal manifold for $\boldsymbol{X}$, if there exists a set set $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ of linear independent principal components for $\boldsymbol{X}$, such that:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{c}_i$$

Then L is termed maximal, iff the sum of the Eigenvalues $\lambda_1, \ldots, \lambda_k$, that correspond to the principal components $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ is maximal.

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

(i) *L minimizes the Euclidean distance to $\boldsymbol{X}$*

(ii) *L is a maximal linear principal manifold for $\boldsymbol{X}$*

Proof. "$\Longrightarrow$" Let $\pi_L \colon \mathbb{R}^d \hookrightarrow L$ denote the orthogonal projection of $\mathbb{R}^d$ onto L. Then according to proposition 3 L minimizes the averaged Euclidean distance to $\boldsymbol{X}$, if and only if (i) $\mathbb{E}(\boldsymbol{X}) \in L$ and (ii) L maximizes the variance $\operatorname{Var}(\pi_L(\boldsymbol{X}))$. In particular (i) is satisfied, if and only if an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ can be chosen, such that:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

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:

$$\operatorname{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{d} \lambda_i a_i$$

Thereupon (ii) is satisfied, if and only if the numbers a_i maximize this sum. Since the covariance matrix $\operatorname{Cov}(\boldsymbol{X})$ is positive semi-definite, the eigenvalues λ_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:

$$\begin{aligned} \sum_{j=1}^{k} \boldsymbol{u}_j^T \operatorname{Cov}(\boldsymbol{X}) \boldsymbol{u}_j &= \sum_{i=1}^{d} \lambda_i a_i \\ &= \sum_{i=1}^{k} \lambda_i \\ &= \sum_{j=1}^{k} \boldsymbol{c}_j^T \operatorname{Cov}(\boldsymbol{X}) \boldsymbol{c}_j \end{aligned}$$

Accordingly the choice $\boldsymbol{u}_j = \boldsymbol{c}_j$ for $i \in \{1, \ldots, k\}$ maximizes $\operatorname{Var}(\pi_L(\boldsymbol{X}))$ and L has a representation, given by:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{c}_i$$

"$\Longleftarrow$" Let L have a representation as given by (ii), then (1) $\mathbb{E}(\boldsymbol{X}) \in L$ and (2) the variance $\operatorname{Var}(\pi_L(\boldsymbol{X}))$ is maximized. According to Proposition 3 it follows, that L minimizes the Euclidean distance to $\boldsymbol{X}$. □

Definition (L-Correlation)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be a random vector and L a maximal linear principal manifold for $\boldsymbol{X}$. Then for any $i, j \in \{1, \ldots, d\}$ let the L-Correlation between X_i and X_j be defined by:*

$$\rho^2_{X_i, X_j | L} \coloneqq R_i R_j \tag{3.13}$$

where with the orthogonal projection $\pi_L \colon \mathbb{R}^d \to L$ *for any* $i \in \{1, \dots, d\}$ *the* ***reliability*** *of* X_i *with respect to* L *is given by:*

$$R_i \coloneqq 1 - \frac{\mathrm{Var}_i\left(\boldsymbol{X} - \pi_L(\boldsymbol{X})\right)}{\mathrm{Var}_i(\boldsymbol{X})} \tag{3.14}$$

Proposition 7. *For an elliptically distributed random vector the L-Correlation generalizes the Pearson Correlation to maximal linear principal manifolds.*

Proof. Let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^2$ be an elliptically distributed random vector and L a maximal linear 1-principal manifold for $\boldsymbol{X}$. Then for $i \in \{1, 2\}$ the random error of the variable X_i has a variance:

$$\eta^2_{X_i} = \mathrm{Var}_{X_i}\left(\boldsymbol{X} - \pi_L(\boldsymbol{X})\right)$$

Such that by the definition of the reliability it follows, that:

$$R_i \overset{3.14}{=} 1 - \frac{\eta^2_{X_i}}{\sigma^2_{X_i}}$$

Consequently:

$$\rho^2_{X_i, X_j | L} = \left(1 - \frac{\eta^2_{X_i}}{\sigma^2_{X_i}}\right)\left(1 - \frac{\eta^2_{X_j}}{\sigma^2_{X_j}}\right)$$

With $n \in \mathbb{N}$ i.i.d. realizations $\boldsymbol{x} \in \mathbb{R}^{n \times 2}$ of $\boldsymbol{X}$ 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:

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

□

4. The Riemann-Pearson Correlation

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

$$\underbrace{\mathrm{Var}(\boldsymbol{X})}_{\text{total variance}} = \underbrace{\mathrm{Var}(\pi_L(\boldsymbol{X}))}_{\text{explained variance}} + \underbrace{\mathrm{Var}(\boldsymbol{X} - \pi_L(\boldsymbol{X}))}_{\text{unexplained variance}}$$

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 $\gamma \colon [a, b] \to \mathbb{R}^d$ (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 $\mathcal{M} \subseteq \mathbb{R}^d$ within the observation space $\mathbb{R}^d$, and an "error free" observation can be identified by a random vector $\boldsymbol{X}^\star$, with outcomes on $\mathcal{M}$. Additionally, however, the observation function may

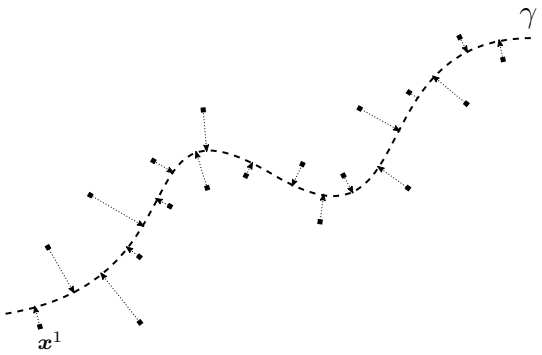

FIGURE 4.1. Principal Curve for a 2-dimensional realization

be regarded to be subjected to a measurement error $\boldsymbol{\varepsilon}$. By the assumption, that $\boldsymbol{\varepsilon}$ has an elliptical distribution, then the distribution of the observable random vector $\boldsymbol{X}$ is represented by a elliptical $\mathcal{M}-$distribution.

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

$$P(\boldsymbol{X}^\star = \boldsymbol{x}) > 0 \Leftrightarrow \boldsymbol{x} \in \mathcal{M} \tag{4.1}$$

Thereupon a random vector $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ is elliptically $\mathcal{M}$-distributed, iff there exists an $\mathcal{M}$-distributed random vector $\boldsymbol{X}^\star \colon \Omega \to \mathbb{R}^d$ and an elliptically distributed random error $\boldsymbol{\varepsilon} \colon \Omega \to \mathbb{R}^d$, such that:

$$\boldsymbol{X} \sim \boldsymbol{X}^\star + \boldsymbol{\varepsilon} \tag{4.2}$$

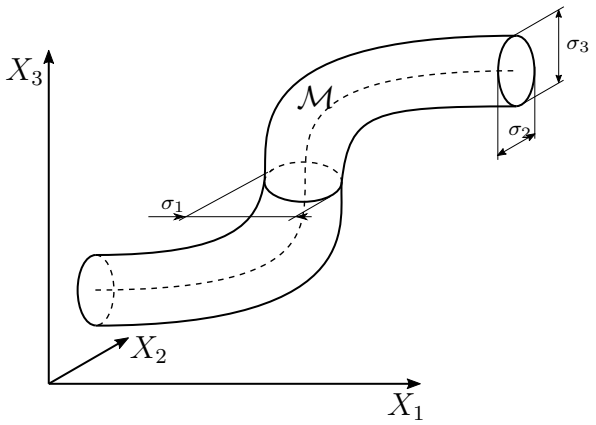

FIGURE 4.2. Elliptical $\mathcal{M}$-distribution in 3 dimensions

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

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

$$\mathbb{E}(\boldsymbol{X} \in \pi_\mathcal{M}^{-1}(\boldsymbol{x})) = \boldsymbol{x} \tag{4.3}$$

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

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 $\boldsymbol{X}$. Intuitively this can be understood as follows: The principal manifold property assures, that:

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(\pi_\mathcal{M}(\boldsymbol{X})) + \mathrm{Var}(\boldsymbol{X} - \pi_\mathcal{M}(\boldsymbol{X}))$$

Consequently the choice of $\mathcal{M}$ maximizes $\mathrm{Var}(\pi_\mathcal{M}(\boldsymbol{X}))$ if and only if it minimizes $\mathrm{Var}(\boldsymbol{X} - \pi_\mathcal{M}(\boldsymbol{X}))$, 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 structural assumptions have to be

incorporated, ether by a parametric family $\{\boldsymbol{f}_\theta\}_{\theta\in\Theta}$ 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 $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a random vector, $\mathcal{M}$ a smooth principal manifold for $\boldsymbol{X}$, which is maximal "with respect to appropriate restrictions" and $\pi_{\mathcal{M}}\colon \mathbb{R}^d \to \mathcal{M}$ a minimal orthogonal projection. Then for any $i, j \in \{1, \dots, d\}$ the Riemann-Pearson Correlation between X_i and X_j is given by:*

$$\rho^2_{X_i,X_j|\mathcal{M}} \coloneqq R_i R_j \int_{\mathcal{M}} S_{i,j}(\boldsymbol{x}) S_{j,i}(\boldsymbol{x}) \,\mathrm{d}P_{\mathcal{M}} \tag{4.4}$$

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

$$R_i \coloneqq 1 - \frac{\mathrm{Var}_i\left(\boldsymbol{X} - \pi_{\mathcal{M}}(\boldsymbol{X})\right)}{\mathrm{Var}_i(\boldsymbol{X})} \tag{4.5}$$

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

$$S_{i,j}(\boldsymbol{x}) \coloneqq \frac{\partial}{\partial x_j}(\boldsymbol{x} - \pi_{\mathcal{M}}(\boldsymbol{x}))\bigg|_i \tag{4.6}$$

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

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

$$S_{i,j}(\boldsymbol{x}) = \frac{\partial}{\partial x_j}(\boldsymbol{x} - \pi_L(\boldsymbol{x}))\bigg|_i \equiv \beta$$

Furthermore for $c \neq 0$:

$$S_{j,i}(\boldsymbol{x}) = \frac{\partial}{\partial x_i}(\boldsymbol{x} - \pi_L(\boldsymbol{x}))\bigg|_j \equiv \frac{1}{\beta}$$

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

$$\begin{aligned}\rho^2_{X_i,X_j|\mathcal{M}} &= R_i R_j \int_{\mathcal{M}} S_{i,j}(\boldsymbol{x}) S_{j,i}(\boldsymbol{x}) \,\mathrm{d}P_{\mathcal{M}} \\ &= R_i R_j \int_{\mathcal{M}} \mathrm{d}P_{\mathcal{M}} \\ &= R_i R_j = \rho^2_{X_i,X_j|L}\end{aligned}$$

□

References

[1] a N Gorban and a Y Zinovyev. Principal Graphs and Manifolds. *Arxiv preprint arXiv08090490*, page 36, 2008.

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

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

[4] 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.

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

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