Bridging Statistics with Geometry and Mechanics

Vladimir Dragović and Borislav Gajić

The University of Texas at Dallas and MISANU, Belgrade

based on:

V. D, B. Gajić, Orthogonal and Linear Regressions and
Confocal Pencils of Quadrics, arXiv: 2209.01679, Satistical Science, under revision

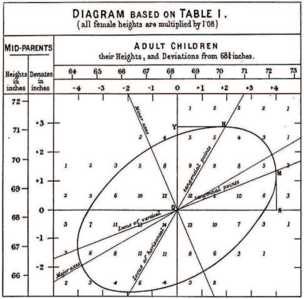
V. D, B. Gajić, Points with rotational ellipsoids of inertia,
envelopes of hyperplanes which equally fit the system of points in $\mathbb{R}^k$...
Physica D: Nonlinear Phenomena, 2023

Third Conference on Nonlinearity, SANS
September 6, 2023, MISANU, Belgrade, Serbia

Francis Galton, 1886, the birth of the law of regression

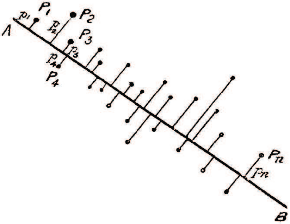
Heredity transmissions-height, 930 children, 205 parents

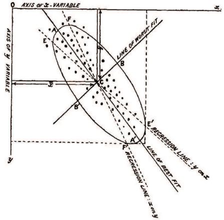
Assigned a “mid-parent” height. Established the average regression from mid-parent to offsprings and from offsprings to mid-parent. Formulated the law of regression toward mediocrity: *When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they. When Mid-Parents are shorter than mediocrity, their Children tend to be taller than they.*



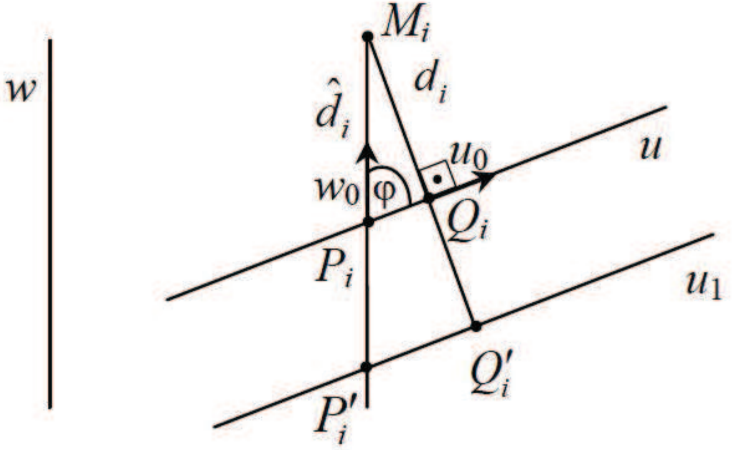
Karl Pearson, 1901, the birth of orthogonal regression

What is the hyper-plane which minimizes the mean square distance from a given set of points in $\mathbb{R}^k$, for any $k \geq 2$? Pearson formulated the problem: *"... we suppose the observed variables–all subject to error–to be plotted in plane, three-dimensioned or higher space, and we endeavour to take a line (or plane) which will be the 'best fit' to such a system of points. Of course the term 'best fit' is really arbitrary; but a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or plane a minimum."*





Classical vs. orthogonal regression



Classical simple regression

It is assumed that the values $(x^{(i)})_{i=1}^N$ are known, fixed values, as for example values set up in advance in the experiment. The values $(y^{(i)})_{i=1}^N$ are observed values of uncorrelated random variables Y_i, $i = 1, \ldots, N$ with the same variance σ^2. A linear relationship is assumed between the predictors $x^{(i)}$ and responses $(y^{(i)})_{i=1}^N$:

$$EY_i = \alpha + \beta x^{(i)}, \quad i = 1, \ldots, N.$$

$$Y_i = \alpha + \beta x^{(i)} + \epsilon_i, \quad i = 1, \ldots, N,$$

ϵ_i are *the random errors* and they are uncorrelated random variables with zero expectation and the same variance σ^2. In such models the regression is of Y on x, i.e. in the vertical direction.

Errors-in-variables (EIV) models, orthogonal regression

Here predictors are known only up to some error. The observed pairs $(x^{(i)}, y^{(i)})_{i=1}^{N}$ are sampled from random variables (X_i, Y_i) with means satisfying the linear relationship

$$EY_i = \alpha + \beta(EX_i), \quad i = 1, \ldots, N.$$

Denoting $EX_i = \xi_i$, the errors in variables model can be defined as

$$Y_i = \alpha + \beta\xi_i + \epsilon_i, \quad X_i = \xi_i + \delta_i, \quad i = 1, \ldots, N,$$

both X_i and Y_i have error terms which belong to mean zero normal distributions, such that all ϵ_i, $i = 1, \ldots, N$ have the same variance σ_ϵ^2 and all δ_i, $i = 1, \ldots, N$ have the same variance σ_δ^2.

There is a symmetry between x_i and y_i as they are both known with an error. It is more natural to apply to them the orthogonal regression, i.e. the orthogonal least square method. We will use it under the assumption that $\eta = \sigma_\epsilon^2/\sigma_\delta^2$ is known and with an analogue assumption in a general dimension k.

Basic definitions

A system of N points $(x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})_{i=1}^N$ is given. *The centroid*, or the mean values of the coordinates $\bar{x}_j$ and *the variances* $\sigma_{x_j}^2$:

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}, \quad \sigma_{x_j}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_j^{(i)} - \bar{x}_j)^2, \quad j = 1, \ldots, k.$$

Due to the generality assumption, all $\sigma_{x_j}^2$, for $j = 1, \ldots, k$ are non-zero. Then, *the correlations* r_{jl} and *the covariances* p_{jl} are

$$r_{jl} = \frac{p_{jl}}{\sigma_{x_j}\sigma_{x_l}},\ p_{jl} = \frac{1}{N}\sum_{i=1}^{N}(x_j^{(i)} - \bar{x}_j)(x_l^{(i)} - \bar{x}_l), \quad j, l = 1, \ldots, k, l \neq j.$$

The covariance matrix K is a $(k \times k)$ with the diagonal elements $K_{jj} = \sigma_{x_j}^2$, and the off-diagonal elements $K_{jl} = p_{jl}$. The covariance matrix is always symmetric positive semidefinite. Here K is positive-definite due to the generality assumption. All its eigenvalues are positive.

The ellipsoid of residuals and Pearson Theorems

Pearson defined *the ellipsoid of residuals* $\sum_{j,l=1}^{k} K_{jl}x_j x_l = const$.
Denote the eigenvalues of K as $\mu_k \geq \cdots \geq \mu_1 > 0$.

Theorem [Pearson]

The minimal mean square distance from a hyperplane to the given set of N points is the minimal eigenvalue of K. The best-fitting hyperplane contains the centroid, it is orthogonal to the corresponding eigenvector of K. It is the principal coordinate hyperplane of the ellipsoid of residuals normal to the major axis.

Confocal families in the k-dimensional space

Family of confocal quadrics

$$Q_\lambda(x_1, \ldots, x_k) : \frac{x_1^2}{\alpha_1 - \lambda} + \cdots + \frac{x_k^2}{\alpha_k - \lambda} = 1.$$

λ – real parameter; $\alpha_1, \ldots, \alpha_k$ – positive real constants;
Jacobi coordinates $(\lambda_1, \ldots, \lambda_k)$: Given $(x_1, \ldots, x_k)$, all λ_j s.t. $Q_{\lambda_j}(x_1, \ldots, x_k)$.

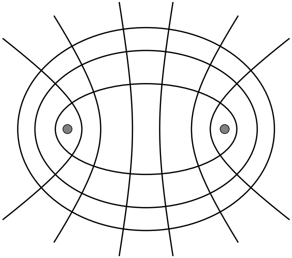
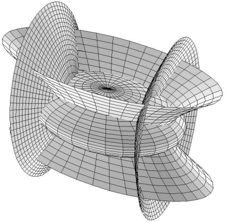


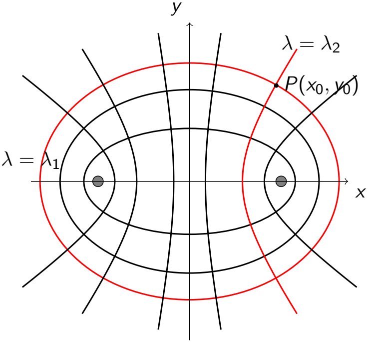
Figure: Confocal family of conics in plane



Confocal conics in plane; Jacobi elliptic coordinates

C. Jacobi, *Lectures on Dynamics* 1842-43, published 1865, Lecture 26

"The main difficulty in integrating a given differential equation lies in introducing convenient variables, which there is no rule for finding. Therefore, we must travel the reverse path and after finding some notable substitution, look for problems to which it can be successfully applied."



Our first goal

For a given system of N points in $\mathbb{R}^k$, for any $k \geq 2$, under the generality assumption, we consider all hyperplanes which equally fit to the given system of points. For any fixed value $s \geq \mu_1$ we consider all hyperplanes for which the mean sum of square distances to the given set of points is equal to s. Starting from the ellipsoid of residuals, we effectively construct a pencil of confocal quadrics with the following property: For each $s \geq \mu_1$ there exists a quadric from the confocal pencil which is the envelope of all the hyperplanes which s-fit to the given system of points.

The ellipsoid of residuals does not belong to the confocal family of quadrics. The construction of this confocal pencil of quadrics is fully effective, though quite involved. The obtained pencil of confocal quadrics is going to have the same center as the ellipsoid of residuals and moreover, the same principal axes.

Our first goal: Example

Example

Let us recall that μ_1 denotes the smallest eigenvalue of the covariance matrix K. In the case $s = \mu_1$ there is only one hyperplane which s fits to the given set of N points. This is the best-fitting hyperplane described in the first Theorem of Pearson. The envelope of this single hyperplane is this hyperplane itself. This hyperplane is going to be a degenerate quadric from our confocal pencil of quadrics.

Our second goal

The second goal:

For a given system of N points in $\mathbb{R}^k$, for any $k \geq 2$, under the generality assumption, find the best fitting hyperplane under the condition that they contain a selected point in $\mathbb{R}^k$. We also provide an answer to the questions of the best fitting line and more general the best fitting affine subspace of dimension ℓ, $1 \leq \ell \leq k-1$ under the condition that they contain a given point.

Our third goal

A careful look at Galton's figure discloses an intriguing geometric fact that the line of linear regression of y on x intersects the ellipse at the points of vertical tangency, while the line of linear regression of x on y intersects the ellipse at the points of horizontal tangency. Further analysis of this phenomenon leads us to our third goal.

The third goal:

To study linear regression in $\mathbb{R}^k$ in an invariant, coordinate free form: *for a given direction and a given system of N points under the generality assumption, what is the best fitting hyperplane in the given direction, among those that contain a selected point in $\mathbb{R}^k$?*

Apparently, the second and the third goal are addressed using the same confocal pencil of quadrics constructed in relation with the first goal and mentioned above.

Hyper-planar moments of inertia

Given a system of points $M_1, ..., M_N$ with masses $m_1, ..., m_N$ in $\mathbb{R}^k$. The hyperplanar moment of inertia for the system of points for a hyperplane π is:

$$J_\pi = \sum_{i=1}^{N} m_i d_i^2. \tag{1}$$

The hyperplanar operator of inertia *at a point O* is:

$$\langle J_O \mathbf{n_1}, \mathbf{n_2} \rangle = \sum_{j=1}^{N} m_j \langle \mathbf{r_j}, \mathbf{n_1} \rangle \langle \mathbf{r_j}, \mathbf{n_2} \rangle, \tag{2}$$

where $\mathbf{r_j}$ is the radius vector of the point M_j; $J_\pi = \langle J_O \mathbf{n}, \mathbf{n} \rangle$, $\mathbf{n}$ is the unit vector orthogonal to $\pi \ni O$. The hyper-planar ellipsoid of inertia *at the point O* is the ellipsoid

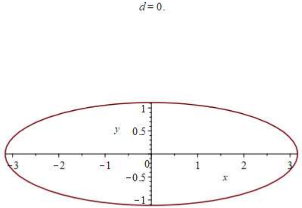
$$\langle J_O u, u \rangle = 1, \quad u \in \mathbb{R}^k.$$

Points with a circle as the ellipse of inertia, $k = 2$

Huygens-Steiner Theorem

If π_1, π_2 are two parallel hyperplanes (lines for $k = 2$) at the distance d and $C \in \pi_1$, then $J_{\pi_2} = J_{\pi_1} + md^2$.

We observed that for any given system of points, there exists a pair of points F_1, F_2 symmetric w.r.t. C, for which the ellipse of inertia is a circle.



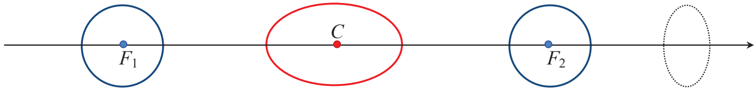
Points with a circle as the ellipse of inertia, $k = 2$

Huygens-Steiner Theorem

If π_1, π_2 are two parallel hyperplanes (lines for $k = 2$) at the distance d and $C \in \pi_1$, then

$$J_{\pi_2} = J_{\pi_1} + md^2.$$

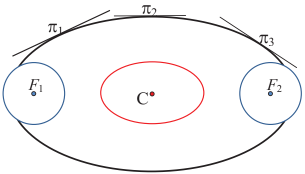
We observed that for any given system of points, there exists a pair of points F_1, F_2 symmetric w.r.t. C, for which the ellipse of inertia is a circle.

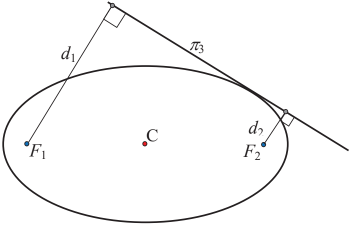


The envelopes of hyperplanes with a given moment, $k = 2$

Theorem: $J_{\pi_1} = J_{\pi_2} = J_{\pi_3}$;

Lemma: $d_1 \cdot d_2 = \text{const.}$

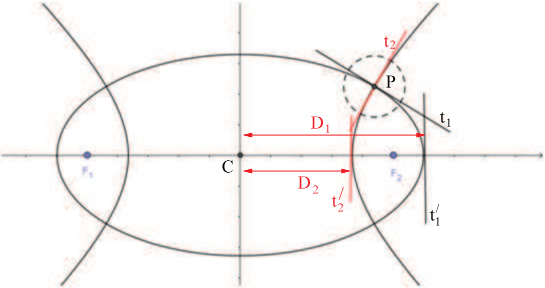




Closing the loop: application to points with rotational ellipsoids of inertia

Proposition [V. D., B. Gajić (2022)]

There are no points with a circle (a rotational ellipsoid) of inertia outside the principal axes.



From a data set to its confocal pencil of quadrics

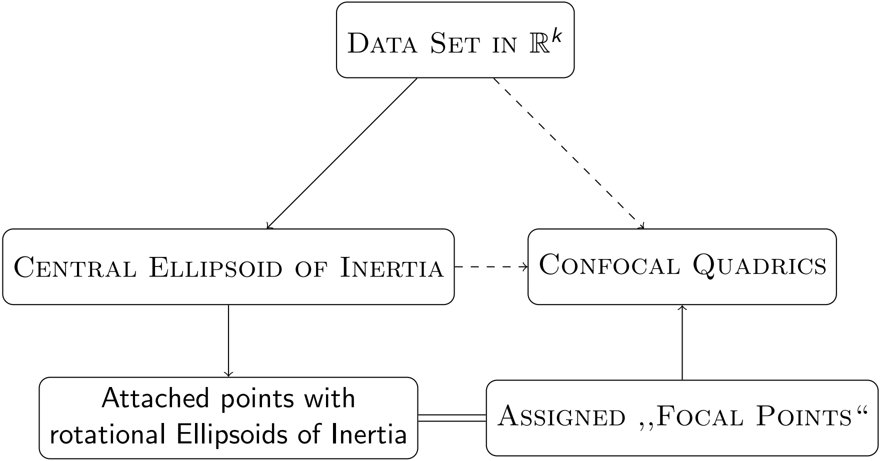


Figure: The construction of the confocal pencil of quadrics starting from a given system of material points in $\mathbb{R}^k$.

The envelopes of hyperplanes with a given moment

Suppose the central principal moments of inertia of a given data set satisfy $0 < J_1 < J_2 < \ldots < J_k$. Let us define $a_2^2, a_3^2, \ldots, a_k^2$ by

$$J_1 + ma_2^2 = J_2, \quad J_1 + ma_3^2 = J_3, \ldots, J_1 + ma_k^2 = J_k.$$

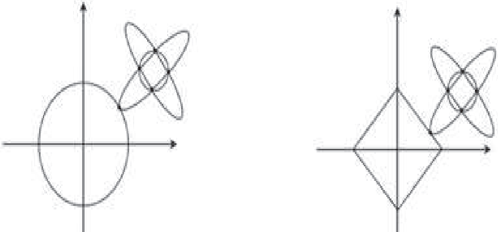
Theorem [V. D., B. Gajić (2022)]

Given a system of points in $\mathbb{R}^k$ of mass m with the central ellipsoid of inertia $J_1x_1^2 + \cdots + J_kx_k^2 = 1$ at the centroid C. The family of hyperplanes for which the system of points has the same hyperplanar moment of inertia are tangent to the same quadric from the pencil of confocal quadrics

$$\frac{x_1^2}{\frac{J_1}{m} - \lambda} + \frac{x_2^2}{\frac{J_1}{m} - a_2^2 - \lambda} + \ldots + \frac{x_k^2}{\frac{J_1}{m} - a_k^2 - \lambda} = 1.$$

Applications to the regularization methods

To study nonlinear constraints on the hyperplanes of regression. May arise in regularization problems for the orthogonal least square method. A ridge-type method imposes an L_2 bound on the coefficients $\beta_1, \ldots, \beta_k$ of the hyperplanes: $||\beta||_2 \leq s$. A lasso-type method assumes use of the L_1 norm and the condition on the coefficients: $||\beta||_1 \leq t$. The best-fit hyperplane under each of these conditions is determined as the point of tangency of a quadric from the linear pencil from Theorem and the L_2 circle of radius s in the first case and the L_1 circle (akka the diamond) of radius t in the second case.



Applications to restricted PCA

Theorem [V. D., B. Gajić (2022)]

Let the points $M_1, ..., M_N$ with masses $m_1, ..., m_N$ be given in $\mathbb{R}^k$ with the associated pencil of confocal quadrics. For any point $P(x_{01}, ..., x_{0k})$, the k tangent hyperplanes to k mutually orthogonal confocal quadrics from the confocal pencil that contain the point P are the principal hyperplanes of inertia at the point P. The obtained principal coordinate axes are the principal components solving Restricted PCA, restricted at the point P, i.e. providing the maximum variance among the normalized combinations $n^T X_P$, uncorrelated with previous ones.

X_P is the $k \times N$ data matrix at the point P, i.e. $J_P = X_P X_P^T$. The eignevalues of J_P vs. the Jacobi coordinates of P:

$$\mu = 2J_1 - m\lambda. \tag{3}$$

$$\Rightarrow J_1 = 2J_1 - m\lambda_{k_C}.$$

Applications to restricted regression

Theorem (generalization of the Pearson Theorem) [V. D., B. Gajić (2022)]

Let the points $M_1, \ldots, M_N$ with masses $m_1, \ldots, m_N$ be given in $\mathbb{R}^k$ with the associated pencil of confocal quadrics. For any point P denote its Jacobi coordinates by $(\lambda_{1_P} < \cdots < \lambda_{k_P})$.

1. The hyperplane of the best fit to the given system of points among the hyperplanes that contain P is the tangent hyperplane to the quadric from confocal pencil with the parameter λ_{k_P}. Similarly, the hyperplane of the worst fit to the given system of points among the hyperplanes that contain P is the tangent hyperplane to the quadric with parameter λ_{1_P}.
2. $\ell : 1 \leq \ell \leq k-1$. The ℓ-moments of π_ℓ and $\hat{\pi}_\ell$:

$$J_{\pi_\ell} = 2(k-\ell)J_1 - m\sum_{j=\ell+1}^{k}\lambda_j, \quad J_{\hat{\pi}_\ell} = 2(k-\ell)J_1 - m\sum_{j=1}^{k-\ell}\lambda_j.$$

Applications to restricted regression

We now illustrate low-dimensional specializations of the last Theorem.

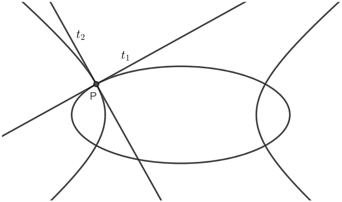


Figure: For $k = 2$: The ellipse and hyperbola from the confocal pencil passing though P. The tangent t_1 to the ellipse at P is the worst fit among all the lines containing P, while t_2, the tangent to the hyperbola at P is the best fit among all such lines. The tangents t_1, t_2 solve RPCA restricted at the point P.

Applications to restricted regression and test statistics

Theorem [V. D., B. Gajić (2022)]

Let the system of N points $M_1, ..., M_N$ with unit masses be given in $\mathbb{R}^k$, $N \geq k$, with the centroid C and the associated pencil of confocal quadrics. For any point P denote its Jacobi coordinates by $(\lambda_{1_P} < \cdots < \lambda_{k_P})$ and of C by $(\lambda_{1_C} < \cdots < \lambda_{k_C})$. Then:

(a) The hyperplanar moment of the best fit is $J_1 = N\lambda_{k_C}$.

(b) The hyperplanar moment of the hyperplane of the best fit that contains the point P is equal to $J_{1_P} = N(2\lambda_{k_C} - \lambda_{k_P})$.

(c) The test statistic of the hypothesis that the hyperplane of the best fit contains the point P is:

$$\frac{N}{N-k+1}(2\lambda_{k_C} - \lambda_{k_P}). \tag{4}$$

whose null distribution can be approximated by Snedecor's F distribution with degrees of freedom $(N-k+1)$ and ∞.

Example: Two types of cells in a fraction of the spleens of fetal mice, the data

Fuller, W. A., *Measurement Error Models*, John Willey and Sons, 1987.

Based on sampling, it is assumed the original counts to be Poisson random variables. The square roots of the counts are given in the last two columns of the table and they have, approximately, constant error variances equal to $T^{-1} = 1/4$, thus $\eta = 1$.

Table: Numbers of two types of cells; following Cohen and D'Eustachio (1978)

j	m_j	n_j	Y_j	X_j
1	52	337	7.211	18.358
2	6	141	2.449	11.874
3	14	177	3.742	13.304
4	5	116	2.236	10.770
5	5	88	2.236	9.381

Example: the model

The model

The postulated model is $y_t = \beta_0 + \beta_1 x_t$,
$(Y_t, X_t) = (y_t, x_t) + (e_t, u_t)$. Y_t is the square root of the number of cells forming rosettes for the t-th individual, and X_t is the square root of the number of nucleated cells for the t-th individual. Based on the sampling, the pair of errors (e_t, u_t) has a covariance matrix, approximately, $\Sigma = T^{-1}E = \mathrm{diag}(0.25, 0.25)$, with $T = 4$.

First calculations

The centroid C is $(\bar{x}, \bar{y}) = (12.7374, 3.5748)$. The components of the hyperplanar inertia operator J_C at C:
$J_{XX} = 47.7937, J_{YY} = 18.1021, J_{XY} = 28.6318$. The principal hyperplanar moments of inertia $J_1 = 0.69605, J_2 = 65.19978$, the eigenvalues of the operator J_C. The corresponding eigenvectors $\mathbf{n}_1 = (-0.51947, 0.85449)^T$, $\mathbf{n}_2 = (-0.85449, -0.51947)^T$ are directions of the principal axes.

Example: the model

The best and worst fit

The line that best fits u_C contains the centroid C and is given by $u_C : y = 0.60793x - 4.16865$. Coincides with Fuller. The equation of the line of the worst fit is $y = -1.64493x + 24.52689$.

Testing hypothesis $\beta_0 = 0$

The origin, denoted by (X, Y), is a point P. $(\tilde{X}, \tilde{Y})$ are the principal coordinates having the centroid C as the origin. The coordinates of P in the principal coordinates: $(\tilde{X}_P, \tilde{Y}_P) = (3.56202, 12.74098)$. The pencil of conics associated with this data: $\alpha = 0.13921$, $\beta = -12.76154$:

$$\frac{\tilde{x}^2}{0.13921 - \lambda} + \frac{\tilde{y}^2}{-12.76154 - \lambda} = 1. \tag{5}$$

The Jacobi elliptic coordinates of the centroid C are: $\lambda_{1_C} = \beta = -12.76154$, $\lambda_{2_C} = \alpha = 0.13921$.

Example: the model

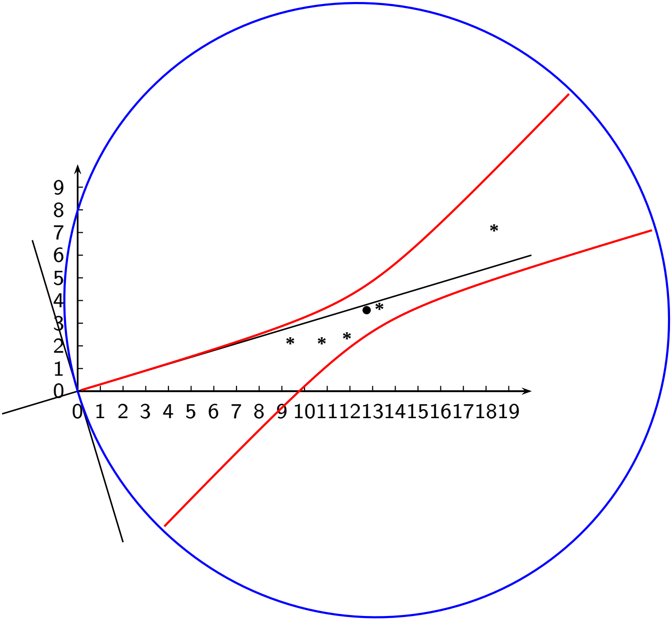
The Jacobi coordinates

From $J_1 = 2J_1 - m\lambda_{2_C}$, we get $J_1 = m\lambda_{2_C} = m\alpha$. The moment of the line u_C is equal to $m\lambda_{2_C} = J_1 = 0.69605$.
The Jacobi elliptic coordinates of P: $\lambda_{1_P} = -186.907$ and $\lambda_{2_P} = -0.73589$.

The line of the best fit at P

The principal moments of inertia at P:
$J_{1_P} = 2J_1 - m\lambda_{2_P} = 5.071564, \quad J_{2_P} = 2J_1 - m\lambda_{1_P} = 935.9271.$
The line u_P the best fit that contains P (in the original coordinates) $u_P : y = 0.30014x$. Thus, $\hat{\beta}_1 = 0.30014$. The moment of u_P is $m(2\lambda_{2_C} - \lambda_{2_P}) = J_{1_P} = 5.071564$, with $m = 5$.

Example: the figure



Example: Testing hypothesis $\beta_0 = 0$

The test statistic

For the null hypothesis $\beta_0 = 0$, the test statistic: $\frac{N}{N-1}\hat{\lambda}$, null distribution can be approximated by Snedecor's F distribution with degrees of freedom $(N-1)$ and ∞. $\hat{\lambda}$ is the smallest root of the equation $\det(J_P - \lambda\Sigma) = 0$.

The test statistic in the Jacobi coordinates

This statistic can be expressed as $\frac{5}{4}\hat{\lambda} = J_{1_P} = 5(2\lambda_{2_C} - \lambda_{2_P})$. The value of J_{1_P} is 5.071564. The approximate p-value is $P(F_{4,\infty} > 5.07) = 0.00043$, which leads to rejection of the null hypothesis.

Thank you!

Acknowledgements

We are grateful to Maxim Arnold, Pankaj Choudhary, Sam Efromovich, and Frank Konietschke for interesting discussions. The research has been partially supported by the Science Fund of Serbia grant Integrability and Extremal Problems in Mechanics, Geometry and Combinatorics, MEGIC, grant no. 7744592 and the Simons Foundation grant no. 854861.

$d = 0.$