Uncertainty Techniques

Statistics, Least Squares, Regression, ML-Estimation, Stochastic Processes

Olaf Wolkenhauer

Control Systems Centre

UMIST

o.wolkenhauer@umist.ac.uk

www.csc.umist.ac.uk/people/wolkenhauer.htm

Contents

1. Learning Objectives

- The expectation operator is a generic concept to summarise information in an underlying universe of discourse.
- Averaging information leads to probability measures and statistics.
- Aggregating information leads to fuzzy measures and possibility measures in particular.
- Matching data to with a model, requires a criterion for how well the data are fitted.
- The least-squares criterion provides optimal parameter estimates for linear models.
- A geometric (vector) representation of the regression problem shows that the optimal solution implies orthogonality.

- The Fourier series is an example for function approximation using the orthogonality principle.
- The least-squares principle does not require a statistical framework to make sense.
- Maximum likelihood estimation is a statistical framework for parameter estimation.
- Stochastic processes are a probabilistic framework to study time-series.
- The Kalman-Bucy filter is a good example how a probabilistic framework, orthogonality and the expectation operator can be used to develop a new concept to model data.

2. The Expectation Operator

The *expectation operator* is a generic tool for the extraction of information. The expectation of any function h with respect to some function g is defined as

$$E[h(\cdot)] \doteq \int_Y h(y) \cdot g(y) \, \mathrm{d}y \; . \tag{1}$$

The expectation operator may be used in two ways to summarise information:

- ▷ *averaging* data to obtain a single reliable measure in the presence of randomness.
- ▷ *aggregating* information to obtain a consensus between similar pieces of information.

2.1. Example: The Probability of an Event

An *event*, is represented by subset $A \subset Y$. Formally an event A is defined by its *characteristic function* ζ :

$$\zeta_A(y) = \begin{cases} 1 & \text{if } y \in A \ , \\ 0 & \text{if } y \notin A \ , \end{cases}$$

The expectation of the characteristic function ζ specifying subset A, then defines the *probability* of event A :

$$\begin{aligned} E[\zeta_A] &= \int_{-\infty}^{+\infty} \zeta_A(y)\ p(y)\ \mathrm{d}y \quad \text{where} \\ &= \int_A p(y)\ \mathrm{d}y \\ &\doteq Pr(A) \ . \end{aligned}$$

2.2. Example: The Probability of a Fuzzy Event

A *fuzzy event* is represented by a *fuzzy set* $A = \{(y, \mu_A(y))\}$, defined by its *membership function* :

$$\begin{aligned} \mu_A : \; Y &\rightarrow [0,1] \\ y &\mapsto \mu_A(y) \end{aligned}$$

The probability of the fuzzy event A is then defined as the expectation of μ_A :

$$\begin{aligned} E[\mu_A] &= \int \mu_A(y) \; \mathrm{d}Pr \\ &= \int_{-\infty}^{+\infty} \mu_A(y) \; p(y) \; \mathrm{d}y \qquad (2) \\ &\doteq Pr(A) \; . \end{aligned}$$

Equation (2) evaluates the degree with which space Y has the fuzzy property A.

2.3. Example: Mean Value and Standard Deviation

Considering values in Y as the outcome of a *random variable* $\mathbf{y}$, a measure of *central tendency* is defined by

$$E[\mathbf{y}] = \int_Y y \cdot p(y)\ \mathrm{d}y$$
$$\doteq \eta\ . \tag{3}$$

η is called the *mean value* of random variable $\mathbf{y}$.

From (3), the dispersion of data in Y, around η, is quantified by the *variance*

$$E[(y-\eta)^2] = \int_Y (y-\eta)^2 \cdot p(y)\ \mathrm{d}y$$
$$\doteq \sigma_{\mathbf{y}}^2\ . \tag{4}$$

The square root of (4) is called *standard deviation*.

2.4. Example: Covariance and Correlation

From (4), considering two random variables $\mathbf{x}$ and $\mathbf{y}$ we define the *covariance* between the two variables as

$$\sigma_{\mathbf{x},\mathbf{y}} \doteq E\big[(\mathbf{x} - \eta_{\mathbf{x}})(\mathbf{y} - \eta_{\mathbf{y}})\big] \ .$$

If $\sigma_{\mathbf{x},\mathbf{y}} = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are said to be 'independent'. A bounded measure of this is the *correlation coefficient* :

$$\rho_{\mathbf{x},\mathbf{y}} \doteq \frac{\sigma_{\mathbf{x},\mathbf{y}}}{\sigma_{\mathbf{x}} \cdot \sigma_{\mathbf{y}}} \qquad \text{where} \qquad -1 \leq \rho \leq 1 \ . \tag{5}$$

A probabilistic model does not refer to a set of sampled data. How do we *estimate* the statistics introduced above...?

2.5. Descriptive Statistics

Given a finite set of data, $\mathbf{M} = \{\mathbf{m}_j \doteq x_j\}$, $j = 1, \ldots, d$, we may for example use the following estimators of (3) and (4) :

$$\hat{\eta} = \frac{1}{d} \sum_{j=1}^{d} x_j \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{d} \sum_{j=1}^{d} (x_j - \hat{\eta})^2 \tag{7}$$

or the *unbiased estimator*

$$\hat{\sigma}^2 = \frac{1}{d-1} \sum_{j=1}^{d} (x_j - \hat{\eta})^2 \ .$$

3. The Least Squares Criterion

The most commonly used criterion to quantify the quality of the model fitting the data is called *least-squared criterion*.

Let

$$\mathbf{x} \doteq [x_1, x_2 \dots, x_r]^T$$

be the *regression vector* over some domain

$$X = (X_1 \times \cdots \times X_r) \subset \mathbb{R}^r \ ,$$

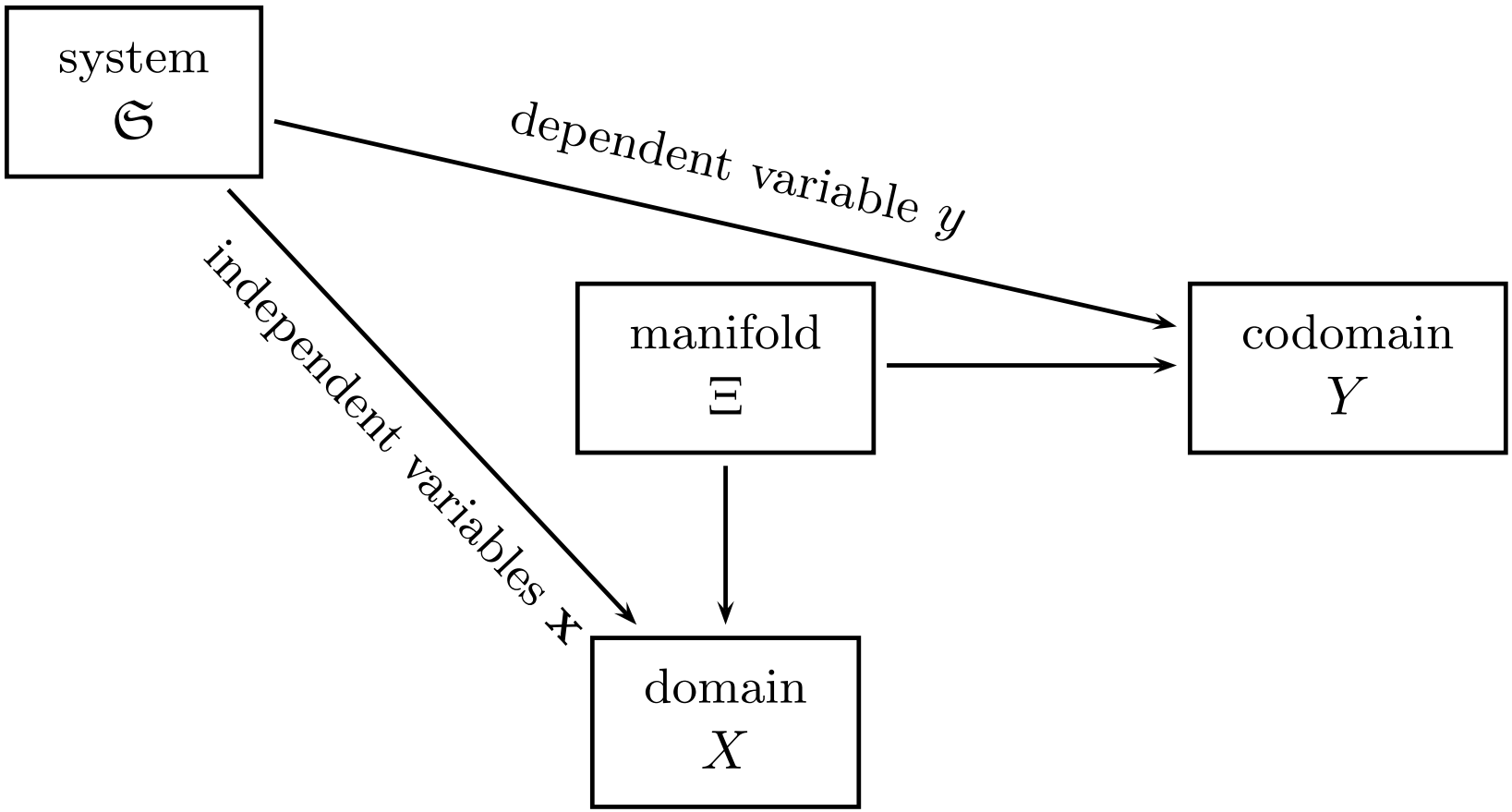
called the *regressor space*.

The aim is then to identify the static dependence

$$y = f(\mathbf{x})$$

of a dependent or *response variable*, $y \in Y \subset \mathbb{R}$, called the *regressand* on the *independent variables* x, called *regressors*.

The aim of the *identification* algorithm is to construct a function $f(\mathbf{x};\boldsymbol{\theta})$, from a finite set of data $\mathbf{M} = \{\mathbf{m}_j\}$, such that $y \approx f(\mathbf{x};\boldsymbol{\theta})$.



3.1. Example: Autoregressive Dynamic Systems

Considering input-output models, using an *auto-regressive* model structure, the system is described by a finite number of past inputs and outputs :

$$\mathbf{x} \doteq [y(k), \ldots, y(k - n_y + 1), u(k), \ldots, u(k - n_u + 1)]^T \ .$$

Linear parametric regression, using the least-squares criterion, provides solutions for linear functions $f(\cdot)$ as discussed in conventional *system identification* [3].

Note that y and $\mathbf{x}$ may not be related to time at all. On the other hand, y may depend only on time, $y = f(t)$, or y is dependent on some variables which themselves vary in time.

3.2. General Regression Model

The problem is to find a function of the regressors

$$f(\mathbf{x};\boldsymbol{\theta}) \ ,$$

called *regression function*, such that the difference,

$$L\big(y, f(\mathbf{x};\boldsymbol{\theta})\big)$$

called *loss* becomes small so that $y = f(\mathbf{x};\boldsymbol{\theta})$ is a good prediction of y. A common loss function for regression is the squared error (L_2) :

$$L\big(y, f(\mathbf{x};\boldsymbol{\theta})\big) = \big(y - f(\mathbf{x};\boldsymbol{\theta})\big)^2 \ . \tag{8}$$

3.3. The Probabilistic Perspective

If y and $\mathbf{x}$ are described within a stochastic framework, one would minimise the expected value of the loss, called the *risk functional* [1] :

$$E[L] = \int L\big(y, f(\mathbf{x};\boldsymbol{\theta})\big)\; p(\mathbf{x}, y)\; \mathrm{d}\mathbf{x}\mathrm{d}y\;. \tag{9}$$

In this case the function f that minimises (8) is the conditional expectation of y given $x_1, x_2, \ldots, x_r$:

$$f(\mathbf{x};\boldsymbol{\theta}) = E[y|\mathbf{x};\boldsymbol{\theta}]$$

.. as the regression of y on $\mathbf{x}$.

3.4. Linear Parametric Regression

In linear parametric regression, y is fit to a *linear combination* of x :

$$f(\mathbf{x};\boldsymbol{\theta}) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_r x_r \tag{10}$$

with vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_r]^T$; written in vector notation,

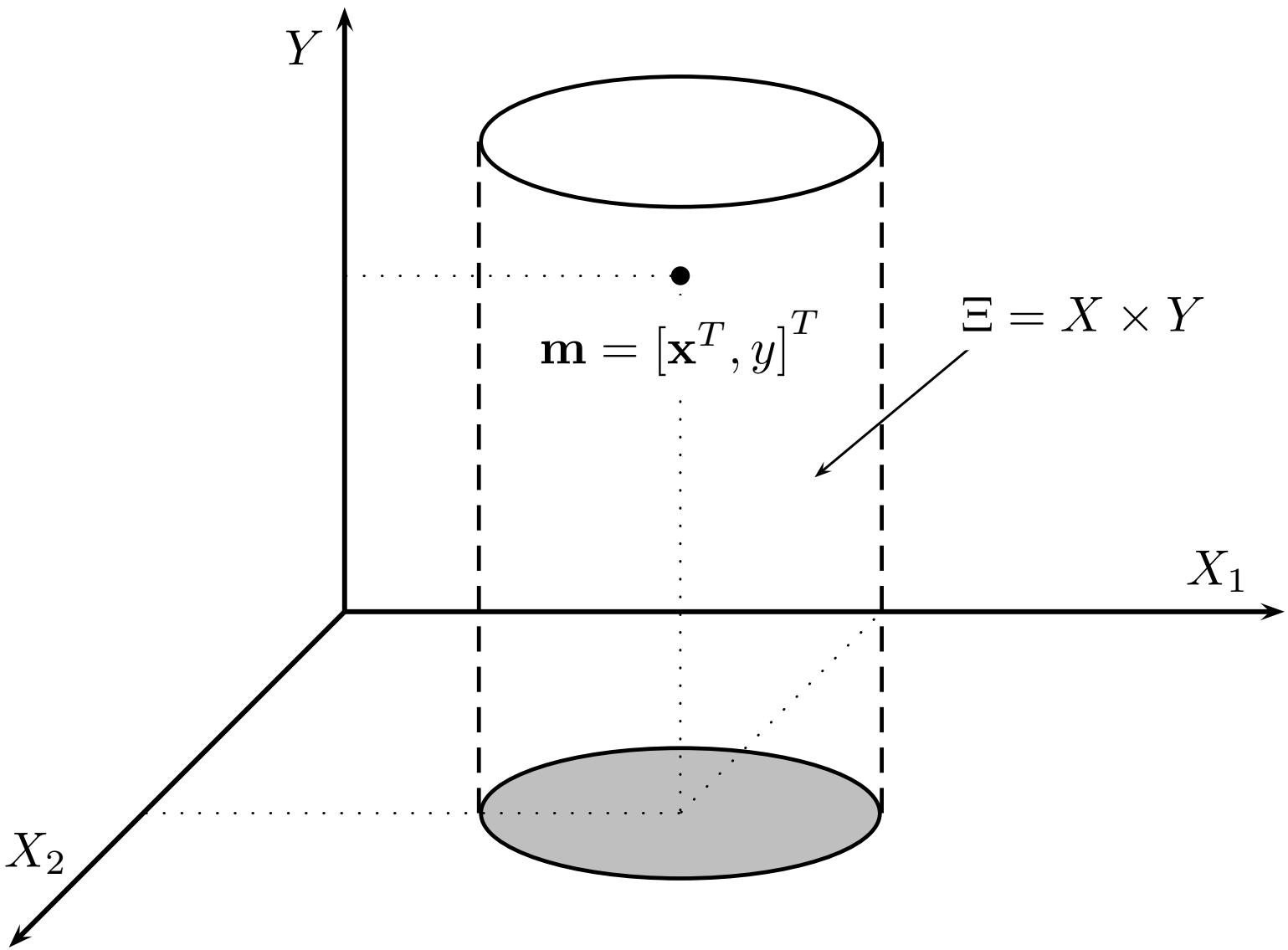
$$f(\mathbf{x};\boldsymbol{\theta}) = \mathbf{x}^T\boldsymbol{\theta} \ .$$

Since only a finite set of sampled data $\mathbf{M} = \{\mathbf{m}_j\}$ is available with

$$\begin{aligned} \mathbf{m}_j &= [\mathbf{x}_j^T, y_j]^T \\ &\doteq [m_{1j}, \ldots, m_{(r+1)j}]^T \in \mathbb{R}^{r+1} \ , \end{aligned} \tag{11}$$

and $[m_{1j}, \ldots, m_{rj}]^T \doteq \mathbf{x}_j \in \mathbb{R}^r$. The variance in (8) is replaced by the sample variance

$$\frac{1}{d}\sum_{j=1}^{d} \big(y_j - f(\mathbf{x}_j)\big)^2 \ .$$


Y
$\mathbf{m} = [\mathbf{x}^T, y]^T$
$\Xi = X \times Y$
X_1
X_2

In the linear case (10), we therefore minimise the variance of the residuals :

$$\hat{\sigma}_e^2 = \frac{1}{d}\sum_{j=1}^{d}\left(y_j - \mathbf{x}_j^T\boldsymbol{\theta}\right)^2 \tag{12}$$

instead of (8). In (12),

$$y - \hat{y} \doteq e \tag{13}$$

is called the *prediction error* which we aim to minimise.

A suitable $\boldsymbol{\theta}$ to choose is the minimising argument of (12) :

$$\hat{\boldsymbol{\theta}} = \arg\min \frac{1}{d}\sum_{j=1}^{d}\left(y_j - \mathbf{x}_j^T\boldsymbol{\theta}\right)^2 , \tag{14}$$

called the *least squares estimate*.

3.5. Derivation of the Solution

Since the loss (12) is a quadratic function of $\boldsymbol{\theta}$, it can be minimised analytically. The necessary condition for the minimum of (14) is, that all derivatives with respect to the parameters θ vanish :

$$\frac{\partial \hat{\sigma}_e^2}{\partial \theta_1} = 2 \sum_{j=1}^{d} x_1 \cdot \big(\theta_1 x_1 + \cdots + \theta_r x_r - y_j\big) = 0$$

$$\frac{\partial \hat{\sigma}_e^2}{\partial \theta_2} = 2 \sum_{j=1}^{d} x_2 \cdot \big(\theta_1 x_1 + \cdots + \theta_r x_r - y_j\big) = 0$$

$$\vdots$$

$$\frac{\partial \hat{\sigma}_e^2}{\partial \theta_r} = 2 \sum_{j=1}^{d} x_r \cdot \big(\theta_1 x_1 + \cdots + \theta_r x_r - y_j\big) = 0$$

These conditions can be rewritten in the form of so-called *normal equations* :

$$\begin{aligned}
\theta_1 \sum x_1 \cdot x_1 + \cdots \theta_d \sum x_1 \cdot x_d &= \sum y_j \cdot x_1 \\
\theta_1 \sum x_2 \cdot x_1 + \cdots \theta_d \sum x_2 \cdot x_d &= \sum y_j \cdot x_1 \\
&\vdots \\
\theta_1 \sum x_d \cdot x_1 + \cdots \theta_d \sum x_d \cdot x_d &= \sum y_j \cdot x_1
\end{aligned}$$

That is, we find that all $\hat{\boldsymbol{\theta}}$ that satisfy

$$\left[\frac{1}{d}\sum_{j=1}^{d}\mathbf{x}_j\mathbf{x}_j^T\right]\hat{\boldsymbol{\theta}} = \frac{1}{d}\sum_{j=1}^{d}\mathbf{x}_j y_j \tag{15}$$

yield a global minimum of (14). If the matrix on the left is invertible, we have

$$\hat{\boldsymbol{\theta}} = \left[\frac{1}{d}\sum_{j=1}^{d}\mathbf{x}_j\mathbf{x}_j^T\right]^{-1} \cdot \frac{1}{d}\sum_{j=1}^{d}\mathbf{x}_j y_j \ . \tag{16}$$

Rewritten in matrix notation, we define the following $d \times 1$ vector and $d \times r$ matrix

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_d^T \end{bmatrix} . \tag{17}$$

The normal equations take the form

$$\left[\mathbf{X}^T\mathbf{X}\right]\hat{\boldsymbol{\theta}} = \mathbf{X}^T\mathbf{Y} \tag{18}$$

and the estimate

$$\hat{\boldsymbol{\theta}} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{Y} \tag{19}$$

where $\left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T$ is known as the Moore-Penrose pseudoinverse and (19) thus gives the solution to the overdetermined ($d > r$) system of linear equations

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} \ . \tag{20}$$

To ensure that $\mathbf{X}^T\mathbf{X}$ is invertible, one needs to choose inputs to the system so that it is "sufficiently excited". If data are to be weighted, we introduce the weighting matrix

$$\mathbf{W} = \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_d \end{bmatrix} \tag{21}$$

and write for (16) and (19),

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \\ &= \left[\sum_{j=1}^{d} w_j \mathbf{x}_j \mathbf{x}_j^T\right]^{-1} \cdot \sum_{j=1}^{d} w_j \mathbf{x}_j y_j \ . \end{aligned} \tag{22}$$

Note that the least-square fitting makes sense without a probabilistic formulation.

3.6. Probabilistic Noise Model

In a stochastic framework, the sequence of regressors $\langle \mathbf{x}(k) \rangle$ is assumed *deterministic*, and the output of the system is a random variable that takes on real values. The output may therefore be interpreted as the sum of a deterministic function and a random error with zero mean, leading to the time-series model

$$y(k+1) = f\big(\mathbf{x}(k); \boldsymbol{\theta}\big) + \varepsilon(k)$$

where $\varepsilon(k)$ is assumed to be a *sequence of independent, identically distributed random variables with zero mean* such that there exists a *population* random variable $\mathbf{y}$ for which

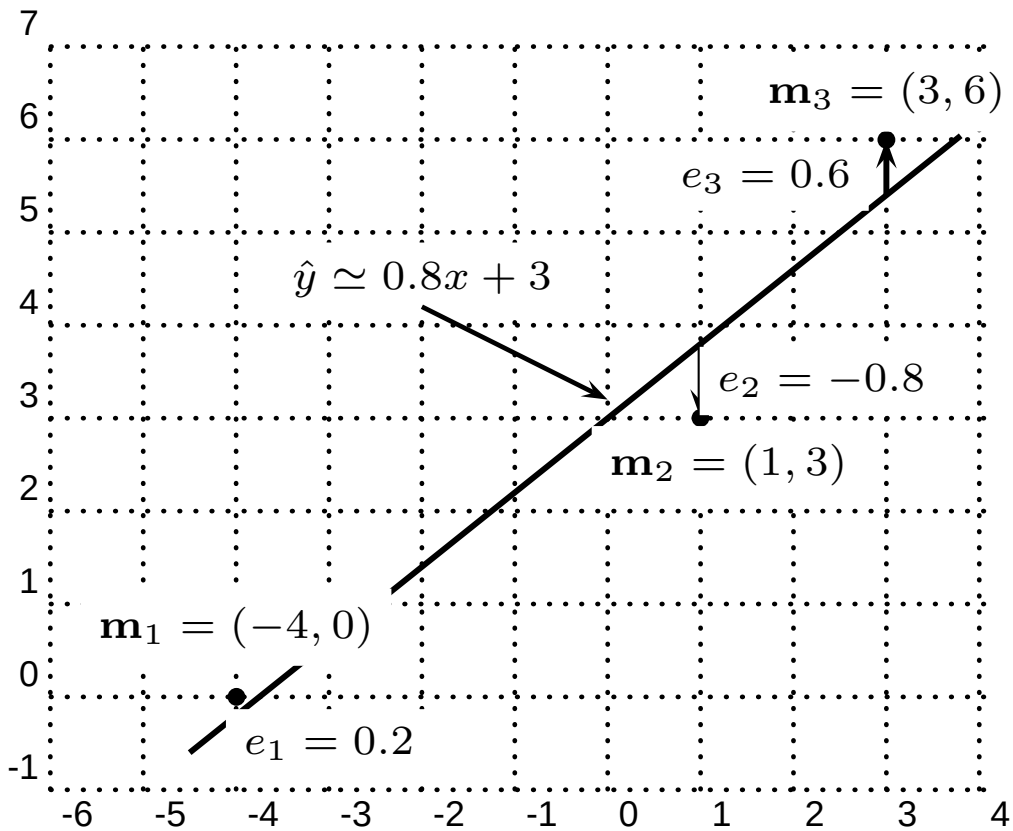
$$E[\mathbf{y}_j] = f(\mathbf{x}_j)$$

and for the *residuals* e_j, $E[e_j] = 0$. Hence the deterministic function is the mean of the output conditional probability

$$f(\mathbf{x}) = \int y \; p(y|\mathbf{x}) \; \mathrm{d}y \; . \tag{23}$$

4. The Geometrical Approach

The unknown *nonlinear* function $y = f(\mathbf{x})$ represents a (non)linear hypersurface in the product space $X \times Y \subset \mathbb{R}^{r+1}$, called *regression surface*. Let us consider three data points $\mathbf{m}_1 = (-4, 0)$, $\mathbf{m}_2 = (1, 3)$, $\mathbf{m}_3 = (3, 6)$ with a regression line fitted through the data.



The model in vector-matrix notation, $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{E}$, where $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$ denotes the residuals, is defined by

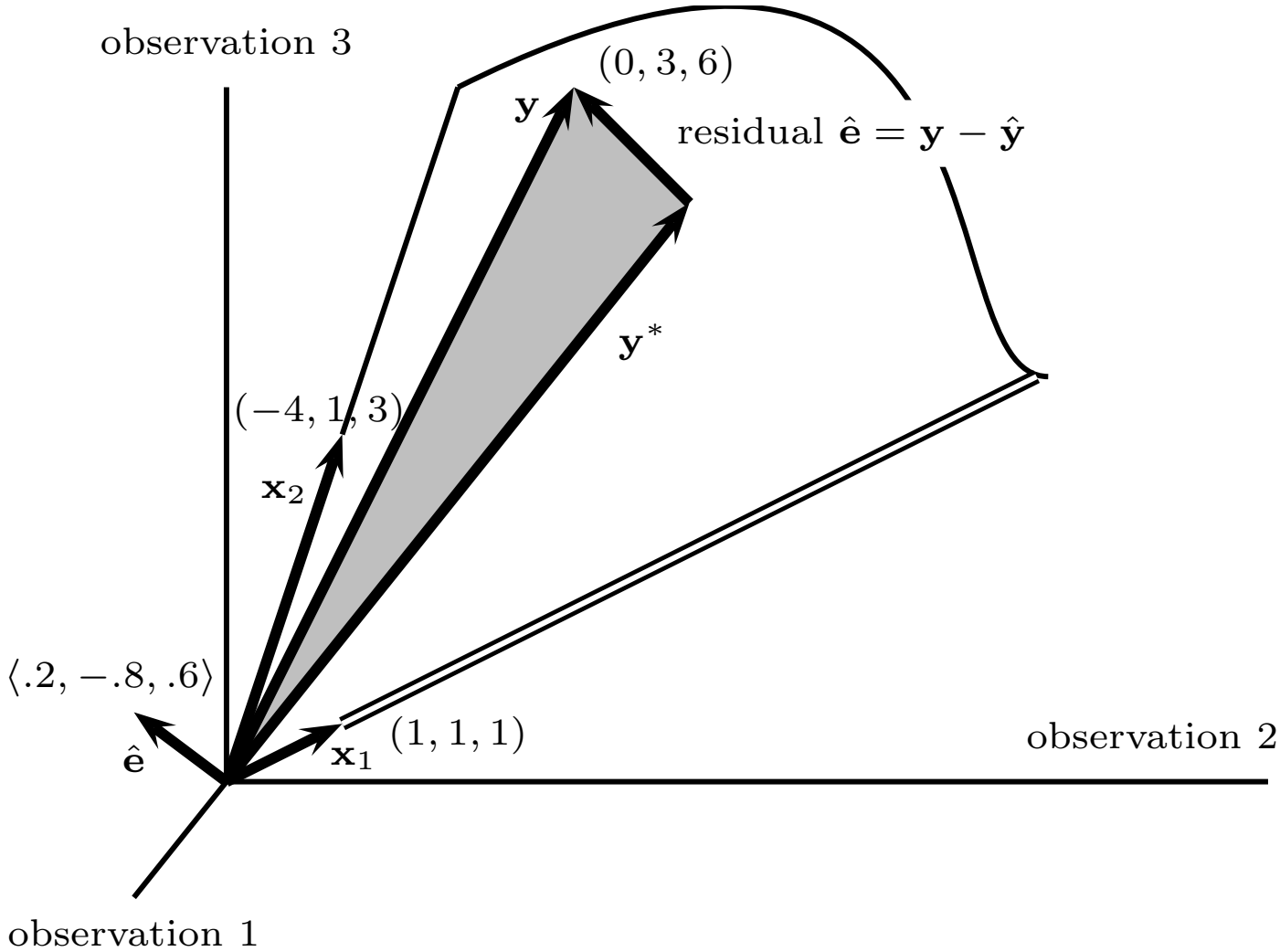
$$\mathbf{Y} = \begin{bmatrix} 1 & -4 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} .$$

The fitted regression line is a vector denoted $\mathbf{y}$. The columns of $\mathbf{X}$ are sequences of sampled values from x_1 and x_2 and are therefore also vectors, denoted by $\mathbf{x}_1$ and $\mathbf{x}_2$,

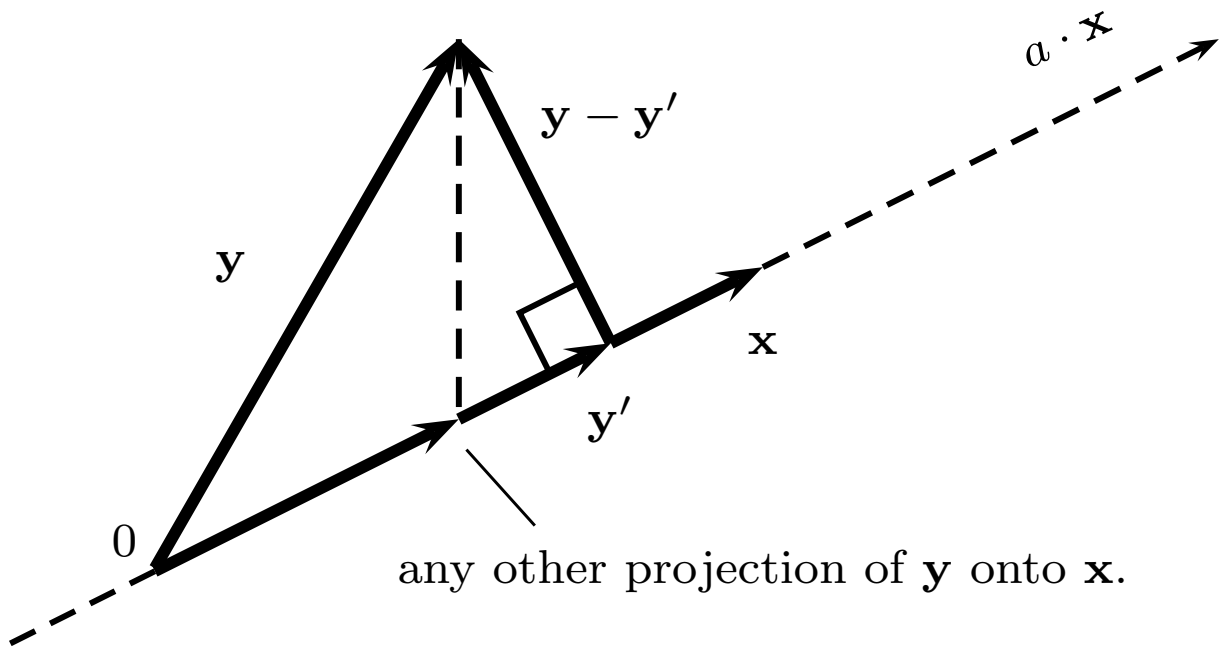
$$\mathbf{y} = \hat{\theta}_1 \cdot \mathbf{x}_1 + \hat{\theta}_2 \cdot \mathbf{x}_2$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \hat{\theta}_1 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \hat{\theta}_2 \cdot \begin{bmatrix} -4 \\ 1 \\ 3 \end{bmatrix} . \tag{24}$$


observation 3
(0, 3, 6)
y
residual ê = y − ŷ
y*
(−4, 1, 3)
x2
⟨.2, −.8, .6⟩
ê
x1
(1, 1, 1)
observation 2
observation 1

Orthogonal projection $\mathbf{y}'$ of $\mathbf{y}$ onto $\mathbf{x}$:



If the regressors are orthogonal, as in figure 1, and we denote by $\mathbf{y}'_1$, $\mathbf{y}'_2$ the perpendicular projections of $\mathbf{y}$ onto $\mathbf{x}_1$ and $\mathbf{x}_2$, we have $\mathbf{y}^* = \mathbf{y}_1{}' + \mathbf{y}'_2$.

Comparing this with (24), we can find simple formulas for θ_1 and θ_2.

For only two vectors $\mathbf{y}$ and $\mathbf{x}$, the perpendicular (orthogonal) projection of $\mathbf{y}$ onto $\mathbf{x}$ is a scalar multiple of $\mathbf{x}$:

$$\mathbf{y}' = a \cdot \mathbf{x} \tag{25}$$

with the problem to determine a such that the inner product $(\mathbf{y} - a \cdot \mathbf{x}) \cdot \mathbf{x} = 0$ is zero, i.e the angle between the two vectors is 90°. Hence,

$$a = \frac{\mathbf{y} \cdot \mathbf{x}}{\mathbf{x} \cdot \mathbf{x}} \ . \tag{26}$$

Substituting (26) into (25), the projection $\mathbf{y}'$ of $\mathbf{y}$ onto $\mathbf{x}$ is

$$\mathbf{y}' = \left(\frac{\mathbf{y} \cdot \mathbf{x}}{\mathbf{x} \cdot \mathbf{x}}\right) \cdot \mathbf{x} \ . \tag{27}$$

From (27), inserted into $\mathbf{y}^* = \mathbf{y}'_1 + \mathbf{y}'_2$, the optimal fit is given as

$$\mathbf{y}' = \left(\frac{\mathbf{y} \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\right) \cdot \mathbf{x}_1 + \left(\frac{\mathbf{y} \cdot \mathbf{x}_2}{\mathbf{x}_2 \cdot \mathbf{x}_2}\right) \cdot \mathbf{x}_2 \ . \tag{28}$$

Comparing (28) with (24), we obtain the parameter estimates as

$$\hat{\theta}_i = \frac{\mathbf{y} \cdot \mathbf{x}_i}{\mathbf{x}_i \cdot \mathbf{x}_i} \ . \tag{29}$$

4.1. Example: Regression Line (Straight Line Fit)

Consider the linear parametric model (10) simplified to

$$\begin{aligned} y &= \theta_1 x_1 + \theta_2 x_2 \\ &\doteq \theta_1 + \theta_2 x \ . \end{aligned} \tag{30}$$

With $x_1 = 1$, we write x for x_2 and use the subscripts for indices of measured values of x. We have the following matrices :

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_d \end{bmatrix} , \qquad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} d & \sum x \\ \sum x & \sum x^2 \end{bmatrix}$$

and for the normal equations (18) and the parameter estimate (19) :

$$\begin{bmatrix} d & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \cdot \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$\hat{\boldsymbol{\theta}} = \frac{1}{d\sum x^2 - (\sum x)^2} \cdot \begin{bmatrix} \sum x^2 \sum y - \sum x \sum xy \\ d\sum xy - \sum x \sum y \end{bmatrix}$$

Considering the two variables y and x, (30) defines a straight line fit through the scatter plot of data for y and x. The slope of the line is determined by parameter θ_2 :

$$\hat{\theta}_2 = \frac{d\sum xy - \sum x \sum y}{d\sum x^2 - (\sum x)^2} \ . \tag{31}$$

The result suggests that the *regression line*, (30) describes also in some way the *correlation* between the two variables x and y. Then there should be a relationship to the correlation coefficient (5) defined earlier.

In a scatter diagram, (see figure 2), the 'cloud' of data is characterised by the estimates of mean values and standard deviation of both variables. More specifically we can draw the σ-line through the point of averages $(\hat{\eta}_x, \hat{\eta}_y)$ with a slope defined by $\hat{\sigma}_y/\hat{\sigma}_x$.

In figure 2 the σ-line and least square fit of the regression line, together with 95% confidence intervals[1], are shown for the following set of data [2] :

x	300	351	355	421	422	434	448	471	490	528
y	2	2.7	2.72	2.69	2.98	3.09	2.71	3.2	2.94	3.73

The point of averages is found at $(\hat{\eta}_x, \hat{\eta}_y) = (422, 2.876)$, $\hat{\sigma}_x^2 = 65.95$, $\hat{\sigma}_y^2 = 0.42$, and $\theta_1 = 0.56, \theta_2 = 0.0055$.

[1] A 95% confidence interval means that we are confident of finding the values in the interval $\pm 2\sigma$ around the regression line. That is, a 95% confidence level means we expect the values to be in the interval 95% of the time.

Replacing the covariance and standard deviations of x and y by their estimators

$$\hat{\sigma}_x = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(x_j - \hat{\eta}_x)^2} \qquad \text{where} \qquad \hat{\eta}_x = \frac{1}{d}\sum_{j=1}^{d} x_j$$

we obtain the following estimate for the correlation coefficient :

$$\hat{\rho}_{x,y} = \frac{\frac{1}{d}\sum(x-\hat{\eta}_x)(y-\hat{\eta}_y)}{\frac{1}{d}\sqrt{\sum(x-\hat{\eta}_x)^2}\sqrt{\sum(y-\hat{\eta}_y)^2}} = \frac{\sum xy - d\hat{\eta}_x\hat{\eta}_y}{\sqrt{\sum(x-\hat{\eta}_x)^2}\sqrt{\sum(y-\hat{\eta}_y)^2}}$$
$$= \frac{\sum xy - d\left[\frac{1}{d}\sum x \cdot \frac{1}{d}\sum y\right]}{\sqrt{\sum(x-\hat{\eta}_x)^2}\sqrt{\sum(y-\hat{\eta}_y)^2}} = \frac{d\sum xy - \sum x \sum y}{d\sqrt{\sum(x-\hat{\eta}_x)^2}\sqrt{\sum(y-\hat{\eta}_y)^2}} .$$

The numerator already matches the one in (31) and we find that if we multiply $\hat{\rho}_{x,y}$ by $\hat{\sigma}_y/\hat{\sigma}_x$, the slope of the σ-line,...

..we find the slope of the regression line coinciding with θ_2 :

$$\begin{aligned}\hat{\rho}_{x,y}\cdot\frac{\hat{\sigma}_y}{\hat{\sigma}_x} &= \frac{d\sum xy-\sum x\sum y}{d\sqrt{\sum\left(x-\hat{\eta}_x\right)^2}\sqrt{\sum\left(y-\hat{\eta}_y\right)^2}}\cdot\frac{\frac{1}{\sqrt{d}}\cdot\sqrt{\sum\left(y-\hat{\eta}_y\right)^2}}{\frac{1}{\sqrt{d}}\cdot\sqrt{\sum\left(x-\hat{\eta}_x\right)^2}}\\ &= \frac{d\sum xy-\sum x\sum y}{d\sum\left(x-\hat{\eta}_x\right)^2} = \frac{d\sum xy-\sum x\sum y}{d\sum x^2-2d^2\hat{\eta}_x^2+d\sum\hat{\eta}_x^2}\\ &= \frac{d\sum xy-\sum x\sum y}{d\sum x^2-d^2\hat{\eta}_x^2} = \frac{d\sum xy-\sum x\sum y}{d\sum x^2-\left(\sum x\right)^2}\\ &= \theta_2\ .\end{aligned}$$

If $\rho_{x,y}$ is exactly $+1$ or -1 we obtain, as a special case, the linear relation $y=\theta_2 x+\theta_1$.

Note, however, variables that are functionally related among each other are correlated but not conversely: if the correlation coefficient is near $+1$ or -1 we may suspect the existence of a law but this is all.

5. Maximum Likelihood Estimation

Let $\mathbf{M} = \{(\mathbf{x}_j, y_j)\}$, denoted $\{\mathbf{m}_j\}$, be a set of d sampled data pairs; the $\mathbf{m}_j$ modelled as outcomes of independent random variables. In the maximum likelihood framework due to R.A Fisher, it is assumed that the data observed a drawn from a distribution with distribution or density

$$p(\mathbf{M}|\boldsymbol{\theta})$$

parametrised by

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_r]^T \ .$$

The key idea in ML estimation is to determine the parameter(s) $\boldsymbol{\theta}$ for which the probability of observing the outcome $\mathbf{M}$ is as high as possible.

The function

$$\ell\left(\boldsymbol{\theta}; \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_d\right) = p(\mathbf{M}|\boldsymbol{\theta}) \tag{32}$$

is called *likelihood function*. The ML estimate of the parameter(s) is that value of parameters which maximises the likelihood function :

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}; \mathbf{M}) \ . \tag{33}$$

Since we maximise ℓ, not the actual value of the function at that point, it is common to ignore constants in the likelihood function that do not depend upon the parameter(s).

In many applications it is more convenient to consider the logarithm of the likelihood function, called the *log-likelihood function* :

$$\mathcal{L}\left(\boldsymbol{\theta}; \mathbf{M}\right) \doteq \ln\ \ell(\boldsymbol{\theta}; \mathbf{M})\ . \tag{34}$$

Since the logarithm is monotonically increasing, maximising the log-likelihood is equivalent to maximising the likelihood.

If the function $\mathcal{L}$ is continuous differentiable, a necessary (but not sufficient) condition to maximise the (log) likelihood is for the gradient to vanish at the value $\boldsymbol{\theta}$ that is the ML value :

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{ML}} \mid \mathbf{M}) = \nabla_{\boldsymbol{\theta}} \ln \mathcal{L}(\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{ML}} \mid \mathbf{M}) = 0 \tag{35}$$

where

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_r}\right]^T .$$

5.1. Example: ML-Estimates for the Normal Distribution

The training data $\mathbf{M} = \{\mathbf{m}_j = x_j\}$ are assumed to derive from the normal distribution

$$p(x;\eta,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\eta)^2}{2\sigma^2}} \quad .$$

The likelihood function takes the form

$$\begin{aligned}\ell(\boldsymbol{\theta};\mathbf{M}) &= p(x_1)\cdot p(x_2)\cdots p(x_d) \\ &= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^d} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{d}(x_j-\eta)^2\right) \quad .\end{aligned}$$

Hence, the log-likelihood function is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta};\mathbf{M}) &= Pr(\mathbf{M}|\eta,\sigma^2) \\ &= -\frac{d}{2}\ \ln(2\pi) - \frac{d}{2}\ \ln(\sigma) - \frac{1}{2\sigma^2}\sum_{j=1}^{d}(x_j-\eta)^2 \ .\end{aligned}$$

We maximise the log-likelihood function by taking the partial derivatives, and equating them with zero

$$\frac{\partial \mathcal{L}}{\partial \eta} = \frac{1}{\sigma^2} \sum_{j=1}^{d} (x_j - \eta) = 0 \tag{36}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{d}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{d} (x_j - \eta)^2 = 0 \ . \tag{37}$$

From (36) and (37) we obtain the ML-estimates as

$$\hat{\eta} = \frac{1}{d} \sum_{j=1}^{d} x_j \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{d} \sum_{j=1}^{d} (x_j - \hat{\eta})^2 \ . \tag{7}$$

5.2. The EM-Algorithm

- For more complicated likelihood functions numerical methods are required for an iterative optimisation.
- A well established example is the *Expectation Maximisation (EM) algorithm*, introduced by A. Dempster.
- The EM algorithm consists of two major steps:
  - ▷ an *expectation step*, followed by a
  - ▷ *maximisation step*.
- The expectation is with respect to the unknown underlying variables, using the current estimate of the parameters and conditioned upon the observations.
- The maximisation step then provides a new estimate of the parameters. These two step are iterated until convergence.

Set loop counter $l = 0$; choose the termination tolerance $\delta > 0$ and initialise parameter(s) $\boldsymbol{\theta}^{(0)}$.

Repeat for $l = 1, 2, \ldots$:

Step 1: **E-Step:** Estimate *unobserved information* using $\boldsymbol{\theta}^{(l-1)}$. The unobserved pdf is

$$p(\mathbf{x}; \boldsymbol{\theta}) \ ,$$

where $\boldsymbol{\theta} \in \Theta$ is the set of parameters of the density. Because we do not have the information of $\mathbf{x}$ to maximise $\ln p(\mathbf{m}; \boldsymbol{\theta})$, we instead maximise the expectation of $\ln p(\mathbf{x}; \boldsymbol{\theta})$ given the data $\mathbf{M}$ and our current estimate of $\boldsymbol{\theta}$:

$$E[\ln p(\mathbf{x}; \boldsymbol{\theta}) | \mathbf{m}, \boldsymbol{\theta}^{(l)}] \doteq Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(l)}) \ .$$

Step 2: **M-Step:** Compute the ML-estimate of parameter(s) $\boldsymbol{\theta}^{(l+1)}$ using information estimated from the E-step :

$$\boldsymbol{\theta}^{(l+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(l)}) \ .$$

Analytically, the ML-estimate is obtained by taking the derivative of $\ln p(\mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, equating it to zero, and solving for $\boldsymbol{\theta}$.

Until $\left\| \boldsymbol{\theta}^{(l)} - \boldsymbol{\theta}^{(l-1)} \right\| < \delta$.

6. Summary

- ✘ The *expectation operator* is a generic tool not restricted to probability theory.
- ✘ (Descriptive) Statistics $\neq$ Probability Theory.
- ✘ The *least squares criterion* makes sense without a probabilistic framework.
- ✘ For a *linear parametric regression model*, we obtain a simple solution from the least squares criterion.
- ✘ In the geometric approach, the optimal least squares estimate corresponds to *orthogonal vectors*.

- The Fourier series is an important example for orthogonal functions and the least squares criterion.
- The *EM-algorithm* is an important tool for maximum likelihood estimation if the distribution of the data is a mixture of density functions.
- A *stochastic process* is a sequence of random variables.
- The Kalman-Bucy filter is a good example how a probabilistic framework, orthogonality and the expectation operator can be used to develop a new concept to model data.

References

[1] Cherkassky, V. and Mulier,F. : *Learning from Data*. Wiley, 1998. 16

[2] Freedman, D. and Pisani, R. and Purves, R. : *Statistics*. Norton, 1997. 34

[3] Ljung, L. : *System Identification*. Prentice Hall, 1987. 14

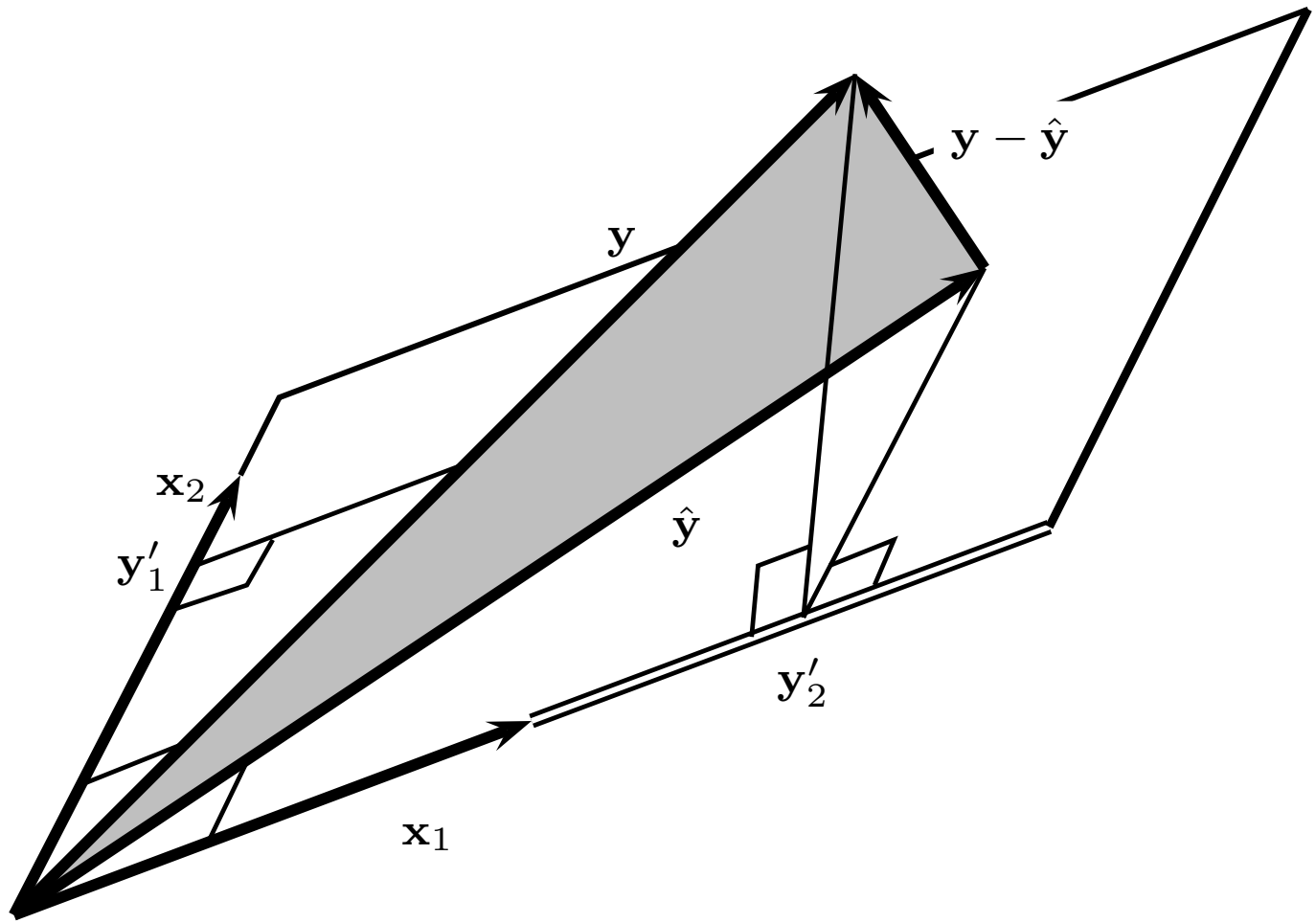


Figure 1: *Vector representation of least-squares regression.*

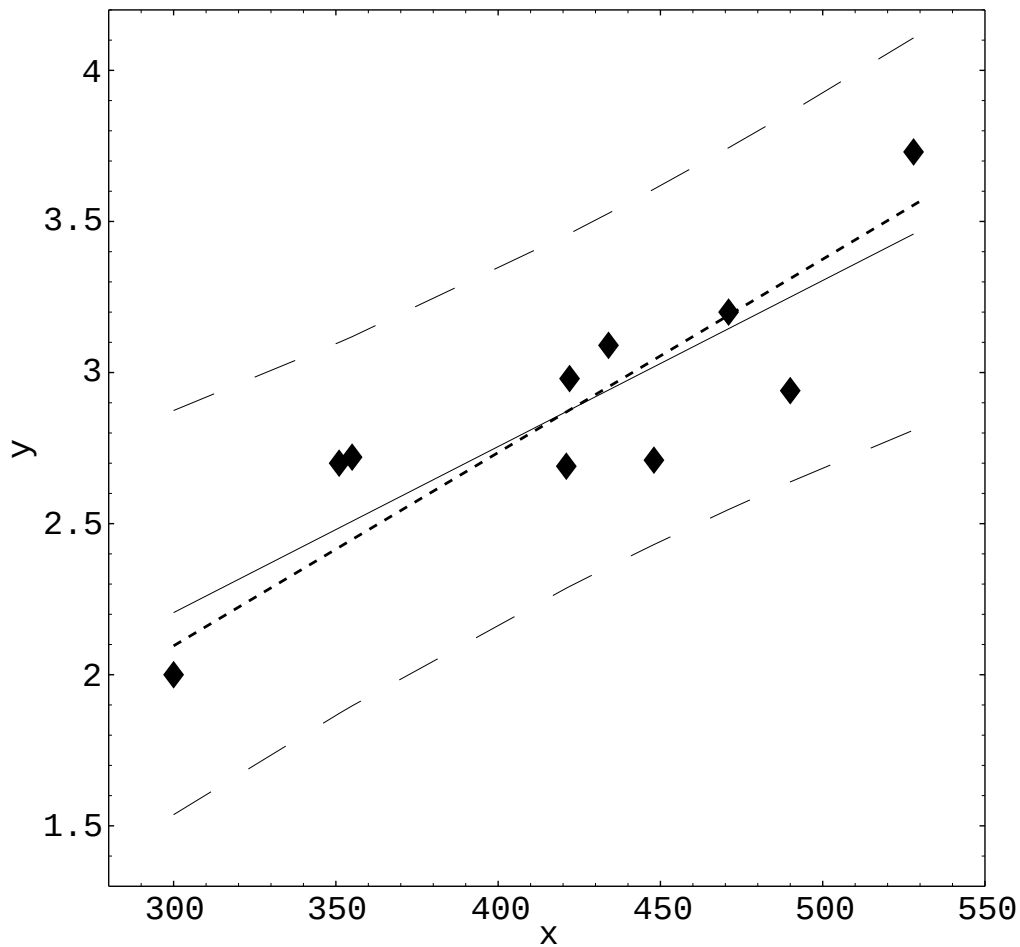


Figure 2: *Regression line (solid), σ-line (dotted), 95% confidence interval (dashed).*