2. Multivariate Statistics

- Geomatics problems normally include the measurement of several quantities. In turn, these measurements are used to determine several unknown parameters.
- The measured quantities cannot usually be treated separately. Instead, they must be dealt with simultaneously. Both the effect of each quantity on the others and the statistical relationship between quantities must be taken into consideration in order to obtain a meaningful solution of the unknowns.
- A multivariate variable consists of more than one different variables, e.g.:

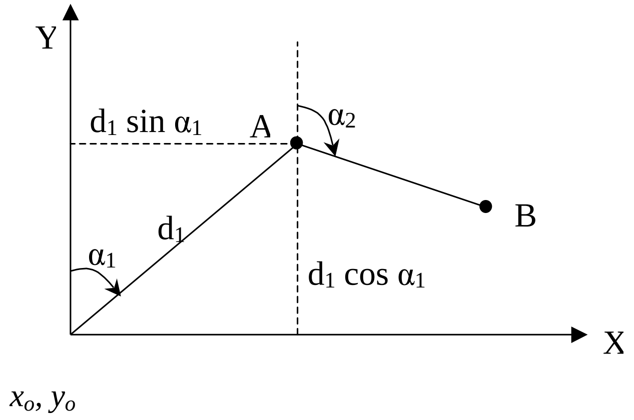
  $\boldsymbol{l} = [\,a, b, c\,]^T$ where a, b, and c, are uni-variables.

  That is,

  $$a = [a_1 , a_2 \ , \ldots a_n]^T \quad \text{with} \quad \bar{a}, \sigma_a, \ and \ \sigma_{\bar{a}}$$
  $$b = [b_1 , b_2 \ , \ldots b_m]^T \quad \text{with} \quad \bar{b}, \sigma_b, \ and \ \sigma_{\bar{b}}$$
  $$c = [c_1 , c_2 \ , \ldots c_k]^T \quad \text{with} \quad \bar{c}, \sigma_c, \ and \ \sigma_{\bar{c}}$$

- Example:

Consider the situation shown in the figure below:



Unknowns

$\boldsymbol{x} = [x_a, y_a, x_b, y_b]^T$

Observations

$\boldsymbol{l} = [\alpha_1 , d_1 , \alpha_2 , d_2]^T$

$$x_a = x_0 + d_1 \sin \alpha_1 \qquad x_b = x_0 + d_1 \sin \alpha_1 + d_2 \sin \alpha_2$$

$$y_a = y_0 + d_1 \cos \alpha_1 \qquad y_b = y_0 + d_1 \cos \alpha_1 + d_2 \cos \alpha_2$$

Any errors in α_1 and d_1 will affect the accuracy of (x_a, y_a)

Any errors in α_1, d_1, α_2 and d_2 will affect the accuracy of (x_b, y_b)

- The following quantities summarises the statistical information of the multivariate variable ***l*** = [a, b, c]^T:

1. The mean of a multivariate variable

$$\bar{\boldsymbol{l}} = \left(\bar{a}, \bar{b}, \bar{c}\right)^T$$

$$\bar{a} = \frac{\sum a_i}{n} \quad \bar{b} = \frac{\sum b_i}{m} \quad \bar{c} = \frac{\sum c_i}{k}$$

2. Variance of the multivariate variable

$$\sigma_l^2 = (\sigma_a^2 \quad \sigma_b^2 \quad \sigma_c^2)^T$$

3. Variance of the mean of the multivariate variable

$$\sigma_{\bar{l}}^2 = (\sigma_{\bar{a}}^2 \quad \sigma_{\bar{b}}^2 \quad \sigma_{\bar{c}}^2)^T = \left(\frac{\sigma_a^2}{n} \quad \frac{\sigma_b^2}{m} \quad \frac{\sigma_c^2}{k}\right)^T$$

2.1. Covariance

- *Covariance* is a measure of the degree of correlation between any two components of a multivariate variable.

- For example, if we have the following set of measurements:

L = (a, b, c)^T where a, b, and c have the **same number of observations**.

Then the covariance between a and b is given by:

$$\sigma_{ab} = \frac{1}{n-1}\sum_{i=1}^{n} v_{ai}\, v_{bi} \qquad \text{(units of a·b),}$$

where σ_{ab} has the same physical meaning as a dot product

$$v_{ai} = \bar{a} - a_i$$

$$v_{bi} = \bar{b} - b_i$$

Note: the number of observations for a and b must be the same

- σ_{ab} has the physical units of *a* multiplied by the physical units of *b*. That is, *the covariance* has no specific units and can take any value between $-\infty \rightarrow +\infty$ (i.e. no limit)

- The covariance between the mean values of *a* and *b* is given by:

$$\sigma_{\bar{a}\bar{b}} = \frac{\sigma_{ab}}{n}$$

- In practice, all the variances and covariances of a multivariate variable are assembled into one matrix called the *variance-covariance matrix* (v-c matrix), or simply the *covariance matrix*.

$$C_l = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 \end{bmatrix}$$

- A similar expression for $C_{\bar{L}}$ can be written for the variance-covariance matrix of the mean.

- If the elements of the multivariate variable are statistically independent (no correlation), then the variance-covariance matrix will be a diagonal matrix.

$$C_l = \begin{bmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_c^2 \end{bmatrix} = diag\left(\sigma_a^2 \quad \sigma_b^2 \quad \sigma_c^2\right)$$

Properties of the variance-covariance matrix

1. Symmetric, that is

$$\sigma_{ij} = \sigma_{ji}$$

2. Its diagonal elements are positive (variances)
3. Non-singular matrix – i.e., the variance-covariance matrix must be invertible. This also means that the determinant of C_l should not equal zero. This property is essential for the purpose of computing the weight matrix P needed in the least squares adjustments, where $P \propto C_l^{-1}$, as will be discussed in later chapters.

- The following matrices cannot be variance-covariance matrices:

$$\mathbf{A} = \begin{bmatrix} 3 & 4 \\ 4 & -1 \end{bmatrix}$$

The variance cannot be a negative number

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & -2 \\ 1 & 3 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

Matrix is not symmetric

$$\mathbf{A} = \begin{bmatrix} 6 & 6 \\ 6 & 6 \end{bmatrix}$$

Matrix is not invertable – determinant is equal to zero (|**A**| = 0)

2.2. Correlation Coefficient

- The correlation coefficient is a measure of how closely two quantities are related to each other.

$$\rho_{ab} = \rho_{ba} = \frac{\sigma_{ab}}{\sigma_a \sigma_b}$$

- Properties of the correlation coefficient

1. Unit-less

eg. $\begin{matrix} a \rightarrow \text{distance } [\text{cm}] \\ b \rightarrow \text{angle } [''] \end{matrix}$ $\sigma_{ab} = [\text{cm} \cdot '']$ $\rho_{ab} = [\text{cm}\cdot'' / (\text{cm}\cdot'')]$

2. Has limits of ± 1

$$\rho_{aa} = \frac{\sigma_{aa}}{\sigma_a \sigma_a} = \frac{\sigma_a^2}{\sigma_a \sigma_a} = 1$$

- If

$\rho_{ab} = 0$	completely uncorrelated
$\rho_{ab} = +1$	completely positively correlated
$\rho_{ab} = -1$	completely negatively correlated
$\lvert\rho_{ab}\rvert < 1$	correlated

$0 < \lvert\rho_{ab}\rvert < 0.35$	weak correlation (i.e. strong solution)
$0.35 < \lvert\rho_{ab}\rvert < 0.75$	significant correlation
$0.75 < \lvert\rho_{ab}\rvert < 1$	strong correlation (i.e. weak solution)

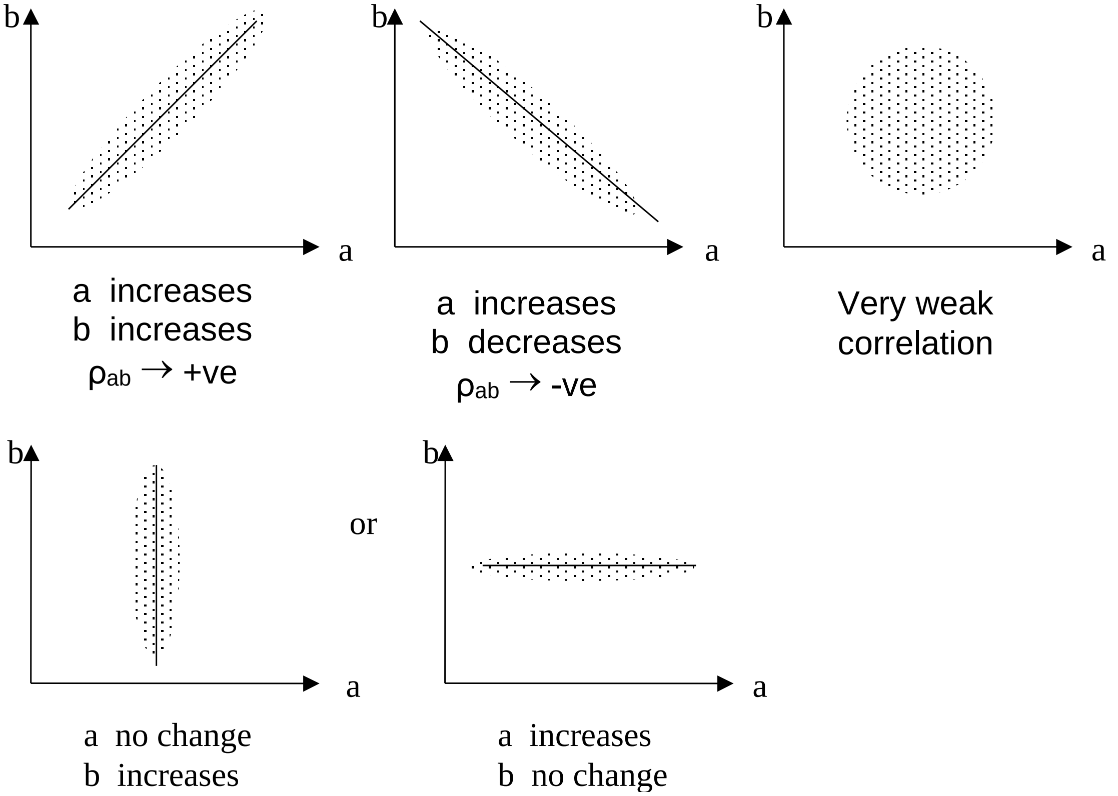
- Similar to the v-c matrix, we can construct the “correlation matrix” $\boldsymbol{\rho}_l$

$$\boldsymbol{\rho}_l = \begin{bmatrix} 1 & \rho_{ab} & \rho_{ac} \\ \rho_{ba} & 1 & \rho_{bc} \\ \rho_{ca} & \rho_{cb} & 1 \end{bmatrix}$$

Note:

$$C_l = \begin{bmatrix} \sigma_a^2 & \rho_{ab}\sigma_a\sigma_b & \rho_{ac}\sigma_a\sigma_c \\ & \sigma_b^2 & \rho_{bc}\sigma_b\sigma_c \\ sym & & \sigma_c^2 \end{bmatrix}$$

2.3. Geometrical Interpretation of the Covariance and Correlation



a increases
b increases
$\rho_{ab} \rightarrow$ +ve

a increases
b decreases
$\rho_{ab} \rightarrow$ -ve

Very weak
correlation

a no change
b increases

a increases
b no change

No correlation

2.4. Mathematical Models

- A mathematical model is comprised of two parts:

1. *Functional Model*: Describes the deterministic (i.e. physical, geometric) relation between quantities

Expresses the functional relationship between quantities

$$\mathbf{f(x,l,c)} = \mathbf{0} \quad (\textit{all may be vector quantities})$$

c...*Constants*

- *e.g. the speed of light*
- *treat as absolute (known) quantities*
- $\sigma_c^2 = 0$ *and* $P \propto {}^{1}\!/_{\sigma_c^2} = \infty$

x...*Unknown parameters*

- *the quantities we wish to solve for*
- *e.g., area of a triangle, co-ordinate* (x, y, z) *of a point*
- *usually treated as having zero weight (but doesn't have to be)*
- $\mathrm{P_x} \propto \dfrac{1}{\sigma_x^2} = 0 \rightarrow \sigma_x^2 = \infty$

l... *Observables*

- *measurements*
- *e.g., distances, angles, satellite pseudoranges*
- $0 < \sigma_l^2 < \infty$

2. *Stochastic Model*: Describes the non-deterministic (probabilistic) behaviour of model quantities, particularly the observations

e.g., $\boldsymbol{l}(a, b)$

$$C_l = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ba} & \sigma_b^2 \end{bmatrix}$$

2.5. Forms of Models

1) Direct Model

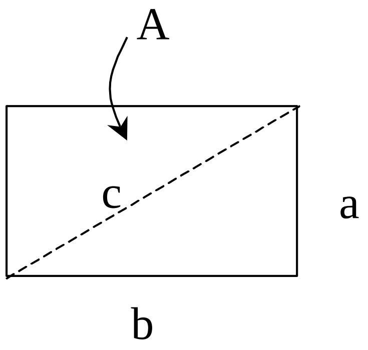
$$\mathbf{x}_{\mathbf{u,1}} = \mathbf{g}_{\mathbf{m,1}}\left(\mathbf{l}_{\mathbf{n,1}}\right)$$

$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_u \end{bmatrix}^T$ *number of unknowns = u*

$\mathbf{g} = \begin{bmatrix} g_1 & g_2 & \dots & g_m \end{bmatrix}^T$ *number of functions = m*

$\mathbf{l} = \begin{bmatrix} l_1 & l_2 & \dots & l_n \end{bmatrix}^T$ *number of observations = n*

- The model is *direct* with respect to the parameters
- One equation per parameter *(i.e. u = m)*
- The parameters are expressed directly as functions of the observations
- Example:



Observations:

$\mathbf{l} = \begin{bmatrix} a & b \end{bmatrix}^T$ *(n = 2)*

Unknowns:

$\mathbf{x} = \begin{bmatrix} A & c \end{bmatrix}^T$ *(u = 2)*

Functions:

$\mathbf{g} = \begin{bmatrix} \mathrm{g}_1 & \mathrm{g}_2 \end{bmatrix}^T$

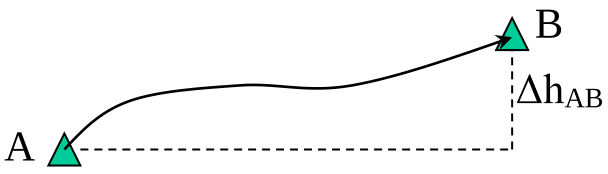
$\mathrm{g}_1 \Rightarrow \mathrm{A} = \mathrm{a} \cdot \mathrm{b}$ *(m = 2)*

$\mathrm{g}_2 \Rightarrow \mathrm{c} = \sqrt{\mathrm{a}^2 + \mathrm{b}^2}$

2) Indirect Model *(Parametric Model)*

$$\mathbf{l}_{\mathbf{n,1}} = \mathbf{h}_{\mathbf{m,1}}\left(\mathbf{x}_{\mathbf{u,1}}\right)$$

- The model is *indirect* with respect to the parameters
- One equation per observation *(n = m)*
- Example: Levelling between two stations (i.e. elevation difference between two stations)



Observations: $l = \Delta h_{AB}$ $(n = 1)$

Unknowns: $x = h_B$ $(u = 1)$

Functions:

$$h = h_1 \qquad (m = 1)$$
$$h_1 \Rightarrow \Delta h_{AB} = h_B - h_A$$

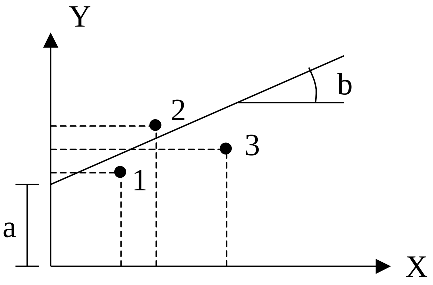
3) Implicit Model *(Combined Model)*

$$\mathbf{f}_{\mathbf{m,1}}(\mathbf{x}_{\mathbf{u,1}}, \mathbf{l}_{\mathbf{n,1}}) = \mathbf{0} \quad (m \neq n \neq u)$$

The model is *implicit* with respect to the parameters and the observations

The parameters and observations cannot be separated, and have an “interwoven” relationship

♦ Example: line fitting



Observations:

$$\mathbf{l} = [x_1 \quad y_1 \quad x_2 \quad y_2 \quad x_3 \quad y_3]^T \quad (n = 6)$$

Unknowns:

$$\mathbf{x} = [a \quad b]^T \quad (u = 2)$$

Functions:

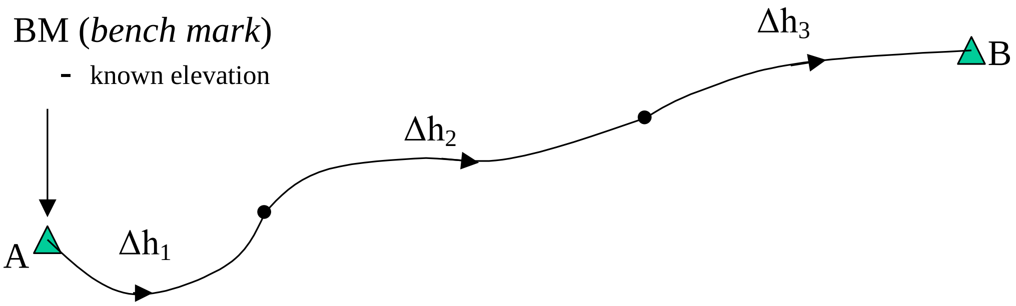
$$\mathbf{f} = [\mathrm{f}_1 \quad \mathrm{f}_2 \quad \mathrm{f}_3]^\mathrm{T} \qquad (m = 3)$$
$$\mathrm{f}_1 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_1 - \mathrm{y}_1$$
$$\mathrm{f}_2 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_2 - \mathrm{y}_2$$
$$\mathrm{f}_3 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_3 - \mathrm{y}_3$$

2.6. Direct Models

$$\boldsymbol{x}_{u,1} = \boldsymbol{g}_{m,1}(\boldsymbol{l}_{n,1})$$ (one equation per parameter)

Direct Linear Models

- The math models are linear w.r.t. to the observations (i.e., when the math models are differentiated, they yield a vector of constants)
- Example – A simple levelling network (note: arrow pointing at the higher station)



▪ Unknowns: $\mathbf{x} = [H_B]$ $u = 1$	▪ Observations: $n = 3$	$\boldsymbol{l} = \begin{bmatrix} \Delta h_1 \\ \Delta h_2 \\ \Delta h_3 \end{bmatrix}$
Constants $\mathbf{c} = [H_A]$	▪ Functions: *m = number of parameters (unknowns)* $m = u = 1$	

Math Model

$$u = 1, m = 1, n = 3 \therefore \textit{unique solution}$$

$$H_B = H_A + \Delta h_1 + \Delta h_2 + \Delta h_3$$

$$[H_B]_{1x1} = [1 \;\; 1 \;\; 1]_{1x3} \begin{bmatrix} \Delta h_1 \\ \Delta h_2 \\ \Delta h_3 \end{bmatrix}_{3x1} + [H_A]_{1x1}$$

Direct Non-linear Models (w.r.t. the observations)

- e.g., $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_1(\mathrm{l}) \\ \boldsymbol{g}_2(\mathrm{l}) \end{bmatrix} = \begin{bmatrix} l_1^2 cos(c) + l_2\ sin(c) \\ \sqrt{l_1 + l_2 + l_3} \end{bmatrix}$

- c is a constant, n = 3, u = m = 2

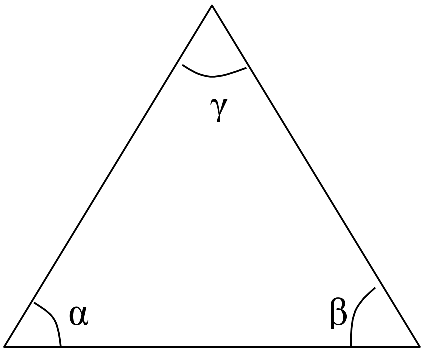
- To solve this problem we usually linearize the model using a *Taylor series expansion (will be discussed in a later chapter)*

Conditional Models

- A special case of the direct model, where no parameters are expressed in the model

$$0 = \boldsymbol{g}_{m,1}\left(\boldsymbol{l}_{n,1}\right)$$

- Example 1 - Estimating the internal angles of a triangle



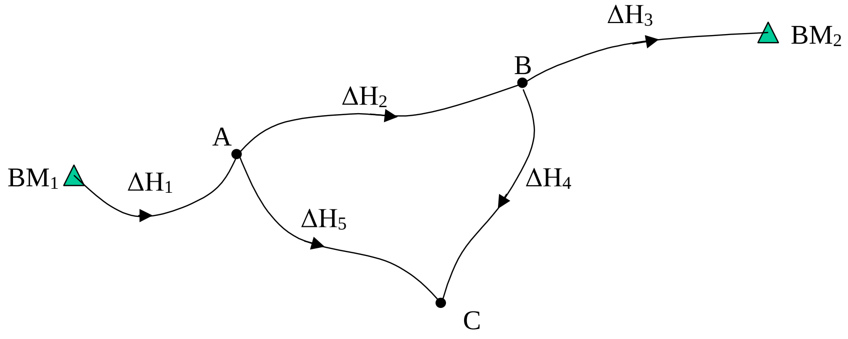
- Unknowns:
  u = 2 (any two angles)
- Observations:
  n = 3
- Functions:
  m = number of independent conditions
  m = n – u = 3 – 2 = 1

Math Model

$$\alpha + \beta + \gamma - 180 = 0$$

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} - 180 = 0$$

♦ Example 2 – A levelling network (note: arrow pointing at the higher station)



▪ Unknowns: $u = 3$	$\mathbf{x} = \begin{bmatrix} H_a \\ H_b \\ H_c \end{bmatrix}$	▪ Observations: $n = 5$	$\mathbf{l} = \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta H_3 \\ \Delta H_4 \\ \Delta H_5 \end{bmatrix}$
Constants	▪ $\mathbf{c} = \begin{bmatrix} H_{BM1} \\ H_{BM2} \end{bmatrix}$	▪ Functions: *m = number of independent conditions* $m = n - u = 5 - 3 = 2$	

Math Model

$$
\begin{aligned}
H_{BM1} + \Delta H_1 + \Delta H_2 + \Delta H_3 - H_{BM2} &= 0 \\
\Delta H_2 + \Delta H_4 - \Delta H_5 &= 0
\end{aligned}
$$

$$
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \end{bmatrix}_{2x5} \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta H_3 \\ \Delta H_4 \\ \Delta H_5 \end{bmatrix}_{5x1} + \begin{bmatrix} H_{BM1} - H_{BM2} \\ 0 \end{bmatrix}_{2x1} = 0
$$

2.7. Indirect Models

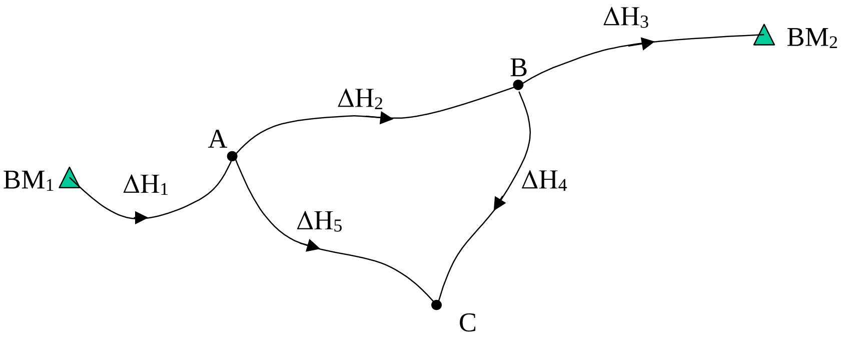
$$\boldsymbol{l}_{n,1} = \boldsymbol{h}_{m,1}\left(\boldsymbol{x}_{u,1}\right) \text{ (one equation per observation)}$$

$$\textit{number of functions} \quad = \quad \textit{number of observations}$$

$$n = m$$

Indirect Linear Models (w.r.t. the parameters)

- Example – Consider the same levelling network (again, arrow pointing at the higher station)



$$\mathbf{x} = \begin{bmatrix} H_a \\ H_b \\ H_c \end{bmatrix}$$

- Unknowns: $u = 3$

$$\mathbf{l} = \begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta H_3 \\ \Delta H_4 \\ \Delta H_5 \end{bmatrix}$$

- Observations: $n = 5$

$$\mathbf{c} = \begin{bmatrix} H_{BM1} \\ H_{BM2} \end{bmatrix}$$

- Constants

- Functions: m = *number of observations*; $m = n = 5$

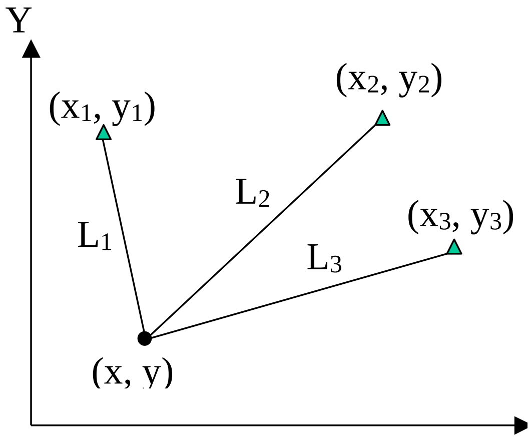
Math Model

$$\begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta H_3 \\ \Delta H_4 \\ \Delta H_5 \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{bmatrix}_{nxu} \begin{bmatrix} H_a \\ H_b \\ H_c \end{bmatrix}_{ux1} + \begin{bmatrix} -H_{BM1} \\ 0 \\ H_{BM2} \\ 0 \\ 0 \end{bmatrix}_{nx1}$$

$$\textit{Redundancy} = m - u = 5 - 3 = 2$$

Indirect Non-linear Models

- ♦ Non-linear models will be linearized using a Taylor series expansion (will be discussed in a later chapter)
- ♦ Example – Finding the co-ordinates of a point by resection



- Unknowns:

$$\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}_{ux1} \quad u = 2$$

- Observations:

$$\mathbf{l} = \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} \quad n = 3$$

- Constants:

$$\mathrm{c} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{y}_1 \\ \mathrm{x}_2 \\ \mathrm{y}_2 \\ \mathrm{x}_3 \\ \mathrm{y}_3 \end{bmatrix}$$

- Functions:

m = number of observations

m = *n* = 3

redundancy = degrees of freedom = m – u = 3 – 2 = 1

Math Model:

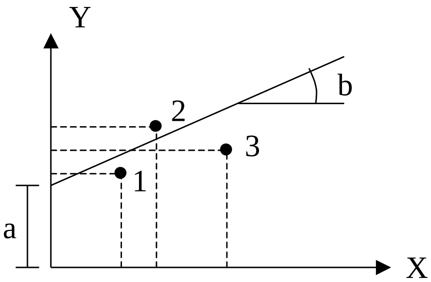
$$\begin{bmatrix} \mathrm{L}_1 \\ \mathrm{L}_2 \\ \mathrm{L}_3 \end{bmatrix} = \begin{bmatrix} \mathrm{h}_1(\mathrm{x}) \\ \mathrm{h}_2(\mathrm{x}) \\ \mathrm{h}_3(\mathrm{x}) \end{bmatrix} = \begin{bmatrix} \left[(\mathrm{x}_1 - \mathrm{x})^2 + (\mathrm{y}_1 - \mathrm{y})^2\right]^{1/2} \\ \left[(\mathrm{x}_2 - \mathrm{x})^2 + (\mathrm{y}_2 - \mathrm{y})^2\right]^{1/2} \\ \left[(\mathrm{x}_3 - \mathrm{x})^2 + (\mathrm{y}_3 - \mathrm{y})^2\right]^{1/2} \end{bmatrix}$$

2.8. Combination of Models

1) Conditions on the Observations

***f*(*x*,*l*) = 0 & *g*(*l*) = 0**

- Example: Consider the example of line fitting and knowing that point (2) is equidistant from points (1) and (3). This condition can be added to the mathematical model



Observations:

$$\mathbf{l} = [x_1 \quad y_1 \quad x_2 \quad y_2 \quad x_3 \quad y_3]^T \quad (n = 6)$$

Unknowns:

$$\mathbf{x} = [a \quad b]^T \quad (u = 2)$$

Functions:

***f*(*x*,*l*) = 0**

$$\mathbf{f} = [\mathrm{f}_1 \quad \mathrm{f}_2 \quad \mathrm{f}_3]^{\mathrm{T}} \qquad (m = 3)$$

$$\mathrm{f}_1 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_1 - \mathrm{y}_1$$

$$\mathrm{f}_2 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_2 - \mathrm{y}_2$$

$$\mathrm{f}_3 \Rightarrow 0 = \mathrm{a} + \mathrm{b}\mathrm{x}_3 - \mathrm{y}_3$$

Adding the condition on the observations:

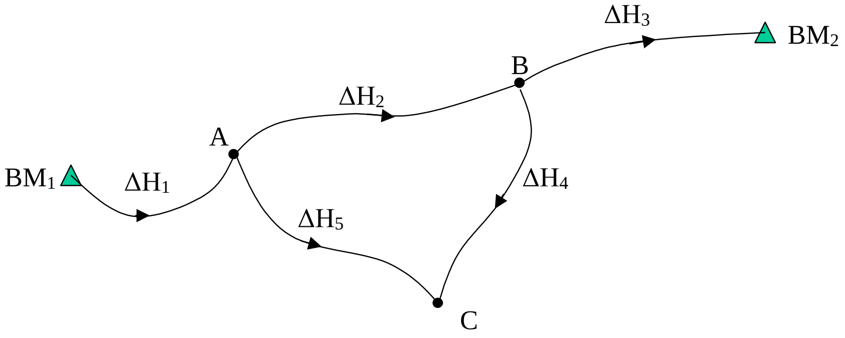
***g*(*l*)** = 0

$$\left[(x_1 - x_2)^2 + (y_1 - y_2)^2\right]^{1/2} - \left[(x_3 - x_2)^2 + (y_3 - y_2)^2\right]^{1/2} = 0$$

2) Conditions on the Unknown Parameters

$\boldsymbol{l} = \boldsymbol{h}(\boldsymbol{x})$ & $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$

- Example – Consider the same levelling network (again, arrow pointing at the higher station) and knowing that stations A and B are at the edge of lake (i.e. $H_a = H_b$)



Math Model

$\boldsymbol{l} = \boldsymbol{h}(\boldsymbol{x})$

$$\begin{bmatrix} \Delta H_1 \\ \Delta H_2 \\ \Delta H_3 \\ \Delta H_4 \\ \Delta H_5 \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{bmatrix}_{nxu} \begin{bmatrix} H_a \\ H_b \\ H_c \end{bmatrix}_{ux1} + \begin{bmatrix} -H_{BM1} \\ 0 \\ H_{BM2} \\ 0 \\ 0 \end{bmatrix}_{nx1}$$

$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$

$H_a - H_b = 0$

$$\begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} H_a \\ H_b \\ H_c \end{bmatrix}_{ux1} = [0]$$