CORRELATIONS FOR ENGINEERING SYSTEMS

by

Dr. Malik Ranasinghe
Senior Lecturer, Department of Civil Engineering
University of Moratuwa

Abstract

Treatment of correlations between variables is necessary to establish realistic modelling of engineering systems. This paper highlights some often ignored theoretical requirements necessary for a rigorous consideration of correlations and presents a general analytical procedure for eliciting correlations between variables which is applicable to a broad class of engineering problems.

The general case of an engineering system described by a decision variable Y, expressed in terms of an arbitrary function $g(\mathbf{X})$ is examined, where $\mathbf{X}$ is a vector of random variables, some or all of which may be correlated. The theoretical requirement for a positive definite correlation matrix is described, along with a two stage process that leads to the elicitation of such a matrix.

A published numerical example is used throughout the paper to demonstrate the approach.

1.0 Introduction

The main objectives in modelling an engineering system are to obtain, a realistic representation of the physical process and a quantification of the uncertainty of that process. Since correlation can have a significant effect on the uncertainty of the decision criteria representing an engineering process, the treatment of correlations between variables is necessary for realistic modelling of engineering systems. Spooner (1974) states that in cost estimating of projects nearly 50% of the total variance can be traced back to co-variance in the labour accounts. Examples highlighting the need for treating correlations in engineering systems include failure functions in structural reliability (Der Kiureghian and Liu, 1986; Foschi et al., 1989), acid rain monitoring stations in water resources (Caselton et al., 1990), estimation of duration, cost and rate of return during the planning stages of an engineering project (Cooper and Chapman, 1987; Perry and Hayes, 1985; Russell and Ranasinghe, 1992).

The objectives of this paper are to highlight some often ignored theoretical requirements necessary for a rigorous consideration of correlations and to present an approach to elicit correlations between variables. The most often suggested analytical approach to model uncertainty in an engineering system is called moment analysis (Benjamin and Cornell, 1970; Ang and Tang, 1975; Bury, 1975; Siddall, 1972; Russell and Ranasinghe, 1992). In this method the uncertainty of the decision variables are quantified as probabilistic moments which are evaluated from probabilistic moments of primary or input variables to the engineering system. The treatment of correlations between primary variables are included in the moment analysis of the decision variables.

This task requires obtaining correlation information, checking for theoretical consistency when this information is elicited and including it in the analysis process. These steps are equally relevant when Monte Carlo simulation (Jaafari, 1988; Perry and Hayes, 1985; Thompson and Wilmer, 1985) or analytical approaches (Cooper and Chapman, 1987; Russell and Ranasinghe, 1992) are used to model an engineering system. For discussion purposes herein, we assume a decision variable Y which represents an engineering system that can be described by an arbitrary function $g(\mathbf{X})$, where $\mathbf{X}$ is the vector of its primary variables, some or all of which may be correlated.

This paper is structured as follows. A numerical example from a published paper in construction engineering is presented in the next section. It is used throughout the paper to illustrate the concepts being discussed. The correlations between primary variables are discussed in the third section. This discussion highlights an often ignored theoretical requirement that the correlation matrix be positive definite. A two stage process which leads to the elicitation of a positive definite correlation matrix for the primary variables is developed in the fourth section. The paper concludes with a step by step procedure for the elicitation of correlations.

2.0 Numerical Example

This numerical example was taken from a recent publication by Russell and Ranasinghe (1992). It describes the duration to fly form a typical slab of 3000 ft^2 in a single suite per floor high-rise building. The duration was considered as the decision variable, and it was estimated from the decomposed relationship given by,

$$T = A + \frac{Q}{P\,L} \qquad (1)$$

in which T was the duration to fly form a typical slab in days, Q was the estimated quantity in ft^2, P was the estimated labour productivity in ft^2*/manhour* once the fly forms were placed, L was the estimated labour usage in *manhours/day* and A was the time required to fly the forms in days. While the quantity was deterministic, the other variables were considered as random. Then, equation re-written as,

$$T = X_1 + \frac{3000}{X_2\,X_3} = g(\mathbf{X}) \qquad (2)$$

was used by Russell and Ranasinghe (1992). The correlation matrix elicited from an experienced engineer and used in their example was as follows.

$$\mathbf{R} = \begin{bmatrix} 1.0 & -0.5 & 0 \\ -0.5 & 1.0 & -0.4 \\ 0 & -0.4 & 1.0 \end{bmatrix}$$

It was stated that the negative correlation between X_1 and X_2 suggests the greater the productivity, probably the greater the efficiency of flying the forms and vice versa. The negative correlation between X_2 and X_3 implies that the smaller the crew the greater the productivity (minimum congestion, all crew members visible and not able to hide).

3.0 Correlation between primary Variables

Generally, when correlations among primary variables are treated it is only the linear correlations. If all the primary variables are normally distributed then the linear correlations between variables are the true correlations. However, in general, all the primary variables are not normally distributed. Consequently, one is faced with the prospect of non-linear correlations. Obtaining and treating non-linear correlations are still complex and largely unresolved theoretical issues. Therefore, we can assume that the correlations among primary variables are linear. This assumption implies that relationships between variables are approximately linear in the region of interest.

In modelling most engineering systems significant data limitations exist. Yet, the treatment of correlations between variables is important to establish realistic quantifications of uncertainty of decision variables. Therefore, the correlations between primary variables may have to be elicited subjectively from experts. A number of authors in both simulation and analytical applications have recognised this necessity (Eilon and Fowkes, 1973; Howard, 1971; Hull, 1977, 1980; Kadane et al., 1980; Keefer and Bodily, 1983; Kryzanowski et al., 1972; Wagle 1967). Except for Kadane et al., (1980), who develop an approach to elicit a positive definite correlation matrix, all of the others obtain only the correlation coefficients.

2.1 Positive Definite Correlation Matrix

A correlation matrix is positive definite if there are no linear dependencies among the primary variables.

Proof that a Correlation Matrix is Positive Definite

Let **X** be the vector of n random variables with covariance matrix $\mathbf{C_x}$ and correlation matrix **R**. Let be a vector of n scalars. From definition,

$$\mathrm{Var}\left[\mathbf{a}^{\mathrm{T}}\,\mathbf{X}\right] \geq 0 \qquad (3)$$

$$\mathbf{a}^{\mathrm{T}}\,\mathbf{C_x}\,\mathbf{a} \geq 0 \qquad (4)$$

Therefore, covariance matrix Cx is always positive positive definite (i.e > 0) or positive semi definite (i.e = 0). Rewriting equation (4) as,

$$\mathbf{a}^{\mathrm{T}}\,\mathbf{C_x}\,\mathbf{a} = \mathbf{a}^{\mathrm{T}}\,(\mathbf{X} - \bar{\mathbf{X}})\,(\mathbf{X} - \bar{\mathbf{X}})^{\mathrm{T}}\,\mathbf{a} \qquad (5)$$

Let $b = (\mathbf{X} - \bar{\mathbf{X}})^T$ a, where b is a number and $b = b^T$. When b = 0 (positive semi definite condition),

$$(\mathbf{X} - \bar{\mathbf{X}})^T \mathbf{a} = 0 \qquad (6)$$

$$(X_1 - \bar{X}_1)\, a_1 + (X_2 - \bar{X}_2)\, a_2 + \ldots\ldots + (X_n - \bar{X}_n)\, a_n = 0 \qquad (7)$$

variable Xn is a linear combination of the others.

If there are no linear dependencies (combinations) then the covariance matrix is always positive definite.

For the correlation matrix, starting from the relationship $\mathbf{R} = \mathbf{D}^{-1}\, \mathbf{C}_X\, \mathbf{D}^{-1}$, where $\mathbf{D}^{-1}$ is the inverse of the diagonal matrix of standard deviations of the **X** vector (i.e **D** = diag[ ]), it follows that,

$$\mathbf{a}^T \mathbf{R}\, \mathbf{a} = \mathbf{a}^T \mathbf{D}^{-1} \mathbf{C}_x \mathbf{D}^{-1} \mathbf{a} \qquad (8)$$

Since $\mathbf{D}^{-1} \equiv [\mathbf{D}^{-1}]^T$ because $\mathbf{D}^{-1}$ is symmetric,

$$\mathbf{a}^T [\mathbf{D}^{-1}]^T \mathbf{C}_x \mathbf{D}^{-1} \mathbf{a} \doteq \mathbf{b}^T \mathbf{C}_x \mathbf{b} \qquad (9)$$

where $\mathbf{b} = \mathbf{D}^{-1} \mathbf{a}$.

Since $\mathbf{D}^{-1}$ is non-singular and symmetric, when $\mathbf{C}_x$ is positive definite,

$$\mathbf{b}^T \mathbf{C}_x \mathbf{b} > 0 \qquad (10)$$

If the covariance matrix is positive definite then the correlation matrix is always positive definite.

If an elicited correlation matrix is not positive definite, then it has to be positive semi-definite because the variance of a vector of random variables is always greater than or equal to zero. This is a theoretical requirement that is ignored in most of the previous developments. A consequence of ignoring this requirement is that the determinant of an arbitrarily generated correlation matrix could be negative and lead to the decision variable having negative variance.

A correlation matrix could be positive semi-definite even when all the variables are not perfectly correlated. For example, suppose the correlation matrix **R** for our three variable example was elicited as follows.

$$\mathbf{R} = \begin{bmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 1.0 & -0.5 \\ 0.5 & -0.5 & 1.0 \end{bmatrix} \qquad (11)$$

At first glance, the correlation coefficients between the variables seem reasonable. However, the determinant of matrix **R** is equal to zero (*i.e.* positive semi-definite). A further investigation shows that a linear combination of variables 2 and 3 is perfectly correlated with variable 1.

4.0 Elicitation of a Correlation Matrix

The proposed method for developing a positive definite correlation matrix consists of a two stage process. The first stage is the elicitation of the linear correlation coefficients among the primary variables, while the second is ensuring the positive definiteness of the correlation matrix.

4.1 Linear Correlation Coefficients

First, the primary variables **X**, in the function for the decision variable $g(\mathbf{X})$, which are considered to be random are ordered according to the expert's confidence in them and their relationship with the other variables. The variable that is selected as the "best" in terms of confidence is numbered one and the "worst" numbered n (when there are n random variables in the function).

The linear correlation coefficient between two primary variables X_i and X_j can then be elicited directly or approximated from the conditional expected value of $X_j|X_i = F$. The conditional expected value of $X_j|X_i = F$ from Bury (1975) is,

$$E[X_j|X_i = F] = E[X_j] + \rho_{ij} \frac{\sigma_j}{\sigma_i} (F - E[X_i]) \qquad (12)$$

Hence,

$$\rho_{ij} = \frac{(E[X_j|X_i = F] - E[X_j])\, \sigma_i}{(F - E[X_i])\, \sigma_j} \qquad (13)$$

in which $E[X_i]$ and $E[X_j]$ are the expected values and σ_i and σ_j are the standard deviations for X_i and X_j; F is the conditional value for X_i; and ρ_{ij} is the linear correlation coefficient between X_i and X_j.

Similar to the method suggested by Hull (1977) for risk simulation, the correlation coefficient between X_i and X_j can be approximated by averaging three or four values for ρ_{ij}. The values for ρ_{ij} are evaluated from equation(13). The conditional expected value of X_j, $E[X_j|X_i = F]$, is elicited by asking the question "What is the anticipated value for X_j, when $X_i = F$?" from experts.

Different percentile values of X_i are used for the conditional value F.

4.2 The Elicitation Procedure

Let $\mathbf{R_n}$ be a subjectively elicited nxn correlation matrix partitioned as,

$$\mathbf{R_n} = \begin{bmatrix} \mathbf{R_{n-1}} & \mathbf{w} \\ \mathbf{w^T} & 1 \end{bmatrix} \qquad (14)$$

where $\mathbf{R_{n\text{-}1}}$ is a $(n - 1) \times (n\text{-}1)$ correlation matrix for $n = 2, 3, \ldots\ldots$ and $\mathrm{wT} = [\rho_{1n}\ \rho_{2n} \cdots\cdots \rho_{n-1n}]$.

Then $\mathbf{R_n}$ is positive definite if $\mathbf{R_{n\text{-}1}}$ is positive definite and,

$$\mathbf{w^T}\ \mathbf{R_{n-1}^{-1}}\ \mathbf{w} < 1. \qquad (15)$$

(Kadane et al., 1980)

If the condition given by equation(15) is violated, then the expert is made aware of the inconsistency and given the option to change one of the correlation coefficient values in the $\mathbf{w}$ vector. When a value is selected (say ρ_{jn}) the expert is informed of the real bounds for ρ_{jn} in which the nxn matrix will be positive definite. These bounds, Γ_1 and Γ_2 (see Fig.1), derived using the condition given by equation(15), are the real roots of the quadratic equation,

$$S_{jj}\,\Gamma^2 + [C_{1j} + C_{2j} + \sum_{i=1}^{j-1} S_{ji}\,W_{1i} + \sum_{i=j+1}^{n-1} S_{ji}\,W_{2i}]\,\Gamma \qquad (16)$$

$$+ \sum_{i=1}^{j-1} (C_{1i} + C_{2i})\,W_{1i} + \sum_{i=j+1}^{n-1} (C_{1i} + C_{2i})\,W_{2i} - 1 < 0$$

where $\mathbf{R_{n-1}^{-1}} = \begin{bmatrix} \mathbf{S_1} \\ \cdots \\ \mathbf{S_j} \\ \cdots \\ \mathbf{S_2} \end{bmatrix}$; $\qquad \mathbf{w^T} = \left[\mathbf{W_1^T} \ \vdots\ \Gamma\ \vdots\ \mathbf{W_2^T} \right]$;

$\mathbf{C_1} = \mathbf{W_1^T}\,\mathbf{S_1}$; and $\mathbf{C_2} = \mathbf{W_2^T}\,\mathbf{S_2}$.
Γ is the correlation coefficient (ρ_{jn}) for which bounds are required,
$\mathbf{S_1}$ is a $(j-1)x(n-1)$ matrix and $\mathbf{S_2}$ is a $(n-1-j)x(n-1)$ matrix,
$\mathbf{W_1^T}$ and $\mathbf{W_2^T}$ are $1x(j-1)$ and $1x(n-1-j)$ row matrices, and
$\mathbf{S_j}$, $\mathbf{C_1}$ and $\mathbf{C_2}$ are $1x(n-1)$ row matrices.

This procedure, of introducing the next ordered variable, eliciting correlation coefficients between that and the previous variables and ensuring that the correlation matrix is positive definite is continued until $\mathbf{R_n}$ is positive definite. Once accepted, the elicitation procedure does not permit the positive definite $\mathbf{R_{n\text{-}1}}$ to be changed.

4.3 Numerical Example

In this section, we use the numerical example to illustrate various aspects of the two stage process just described. The correlation matrix elicited for this example from an experienced engineer was given as,

$$\mathbf{R} = \begin{bmatrix} 1.0 & -0.5 & 0 \\ -0.5 & 1.0 & -0.4 \\ 0 & -0.4 & 1.0 \end{bmatrix}$$

Assume that the experienced engineer had selected A, the time required to fly the forms as the first variable while L, the estimated labour usage had been selected as the third variable. The value elicited directly for ρ_{12} from the expert was equal to -0.5. It was stated that the negative correlation between variables $A(X_1)$ and $P(X_2)$ suggests the greater the productivity, probably the greater the efficiency of flying the forms and vice versa.

Instead of eliciting ρ_{12} directly, if the process suggested above is adopted, a hypothetical set of conditional expected values for X_2, $E[X_2\ X_1 = F]$, that could be given by the expert to get $\rho_{12} = -0.5$ is given in Table 1. The percentile values for A used as F and the ρ_{12} values that are evaluated from equation(13) are also given in Table 1. The percentile values for A assumed as the value for F were taken from Table 1 of Russell and Ranasinghe (1992).

Table 1: Elicitation of ρ_{12}

Percentile	5%	25%	75%	95%
F	0.25	0.33	0.42	0.5
$E[X_2 \mid X_1 = F]$	21.0	20.25	19.4	18.6
ρ_{12}	−0.49	−0.5	−0.48	−0.51

For this numerical example, ρ_{12} is extremely sensitive to the elicited conditional expected value for X_2. For example, suppose that $E[X_2|X_1 = 0.33]$ was given as 20.5 instead of 20.25. From equation (13)

$$\rho_{12} = \frac{(E[X_2|X_1 = 0.33] - E[X_2])\,\sigma_1}{(0.33 - E[X_1])\,\sigma_1} \qquad (17)$$

Since $E[X_1] = 0.375$; $E[X_2] = 19.815$; $= 0.08$ and $= 1.54$ (from Table 2 of Russell and Ranasinghe, 1992), from equation (17) the value for ρ_{12} is equal to -0.79. While this sensitivity cannot be generalised to all situations, it highlights the inherent difficulties one faces when trying to elicit correlation coefficient values. Development of a robust elicitation process is seen as crucial to the practical treatment of correlations. The use of graphics is suggested as being central to this process.

An important part of the elicitation procedure is to have the expert explain and preferably sketch the nature of the relationships between variables, and then identify the region of interest where an estimate of correlation is desired. For the numerical example, an approximation of the relationship between productivity P and labour usage L is shown in Fig.2. It is non-linear, and neither monotonically increasing or decreasing. During the elicitation process, the expert is asked to exercise judgement regarding the practical region of the relationship. Thus, it is explained that below a certain threshold value, crew sizes are too small to permit effective handling of the elements involved because of their size and weight. Once crew sizes get too large, productivity drops off quickly because of congestion and supervision problems. By eliminating small and large crew sizes, the idealised relationship shown in Fig.3 which represents a plot of $P(X_2)$, versus $L(X_3)$ can be sketched. This figure, combined with estimates of $E[X_2]$, $E[X_3]$, σ_2, σ_3 can then be used to estimate σ_{23}. In general this process is imprecise because the relationships between variables are shaped by many uncontrollable factors, and are at best subjective estimates based on extensive experience.

The value elicited directly for ρ_{12} from the expert is considered to be consistent with the expert's belief because the 2x2 matrix is always positive definite. Thereafter, values for $\rho_{13} = 0.0$ (uncorrelated) and $\rho_{23} = -0.4$ respectively are considered. Since the condition given by equation (15) is satisfied for ρ_{13} and ρ_{23}, the correlation values are accepted because the 3x3 matrix is positive definite.

However, if the condition was violated, for example if the expert had said that ρ_{13} was equal to -0.65 instead of 0.0, then the expert is made aware of the inconsistency and given the option to change one of the correlation coefficient values in the w vector. When a value is selected (say ρ_{12}) the expert is informed that the bounds for ρ_{12} in which the 3x3 matrix will be a positive definite correlation matrix from equation (16) are -0.59 and 0.99. Such guidance is given by an interactive computer program called "CORMAT". In evaluating these bounds it is assumed that all of the other correlation coefficient values are consistent with expert belief.

5.0 Summary

We conclude by summarising the developed approach for obtaining correlations in an engineering system.

1. Model the engineering system as a function between a decision variable Y and its primary variable X - *i.e* $g(X)$.

2. Order the primary variables in order of confidence - the variable about which the most is known in terms of statistical relationships with others is first.

3. Elicit the correlation between the first and second variables in the list. If historical data exist, estimate ρ_{ij} directly using standard computation procedures. If no data is available, have the expert explain/sketch the perceived relationship between the variables. Then estimate ρ_{ij} using equation (13).

4. Continue the above procedure by adding one variable at a time while ensuring the positive definiteness of the correlation matrix **Rn**. If **Rn** is not positive definite, use equation(16) to compute acceptable bounds on ρ_{ij} which are fed back to the expert for consideration.

Acknowledgements

The continuing support and facilities provided by the University of Moratuwa to carry out this research is gratefully acknowledged.

References

Ang, A.H-S., and Tang, W.H. (1975), *Probability Concepts in Engineering Planning and Design,* John Wiley & Sons, New York.

Benjamin, J.R., & Cornell, C.A. (1970), *Probability, Statistics and Decision for Civil Engineers*, McGraw-Hill Book Co., New York.

Bury, K.V. (1975), *Statistical Models in Applied Science,* John Wiley & Sons, New York.

Caselton, W.F., Kan, L., and Zidek, J.V. (1990), "Quality Data Networks that Minimize Entropy", *Statistics in the Environmental and Earth Science,* Griffin Books, U.K.

Cooper, D.F., and Chapman, C.B. (1987), *Risk Analysis for Large Projects, Models, Methods and Cases,* John Wiley & Sons, New York.

Der Kiureghian, A., and Liu, P-L. (1986), "Structural Reliability under Incomplete Probability Information", *Journal of the Engineering Mechanics Division, ASCE,* Vol.112, No.1, pp. 85-104.

Eilon, S. and Fowkes, T.R. (1973), "Sampling Procedures for Risk Simulation", *Operational Research Quarterly* Vol.24, No.2, pp. 241-252.

Foschi, R.O., Folz, B.R., and Yao, F.Z. (1989), "Reliability Based Design of Wood Structures", *Structural Report Series, Report #34,* Department of Civil Engineering, The University of British Columbia, Vancouver, Canada.

Howard, R.A. (1971), "Proximal Decision Analysis", *Management Science,* Vol. 17, No.9, pp. 507-541.

Hull, J.C. (1977), "Dealing with Dependence in Risk Simulation", *Operational Research Quarterly,* Vol.28, No.1ii, pp. 201-213.

Hull, J.C. (1980), *The Evaluation of Risk in BusinessInvestment,* Pergamon Press.

Jaafari, A. (1988), "Project Viability and Economic Risk Analysis", *Journal of Management in Engineering, ASCE,* Vol.4, No.1, pp. 29-45.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1980), "Interactive Elicitation of Opinion for a Normal Linear Model", *Journal or American Statistical Association,* T & M Section, Vol.75, pp. 845-854

Keefer, D.L., and Bodily, S.E. (1983), "Three-Point Approximations for Continuous Random Variables", *Management Science,* Vol.29, No.5, pp. 595-609.

Kryzanowski, L., Lusztig, P., and Schwab, B. (1972), "Monte Carlo Simulation and Capital Expenditure Decisions - A Case Study",*Engineering Economist,* Vol.18, No.1, pp. 31-48.

Perry, J.G., and Hayes, R.W. (1985), "Risk and its Management in Construction Projects", *Proceedings, Institution of Civil Engineers,* U.K, Vol.78, No.1, pp. 499-521.

Russell, A.D. & Ranasinghe, M. (1992), "Analytical Approach for Economic Risk Quantification of Large Engineering Projects", *Construction Management and Economics,* E & FN Spon, U.K., Vol. 10, pp.277-301.

Siddall, J.N. (1972), *Analytical Decision Making in Engineering Design,* Prentice - Hall, Englewood Cliffs, N.J.

Spooner, J.E. (1974), "A Mathematical Model for Contingency", *Transactions, American Association of Cost Engineers.* pp. 81-88.

Thompson, P.A. and Wilmer, G. (1985), "CASPAR - A Program for Engineering Project Appraisal and Management", *2nd International Conference in Civil & Structural Engineering Computing,* Vol. 1, pp. 75-81.

Wagle, B. (1967), " A Statistical Analysis of Risk in Capital Investment Projects ", *Operational Research Quarterly,* Vol.18, No.1, pp. 13-33.

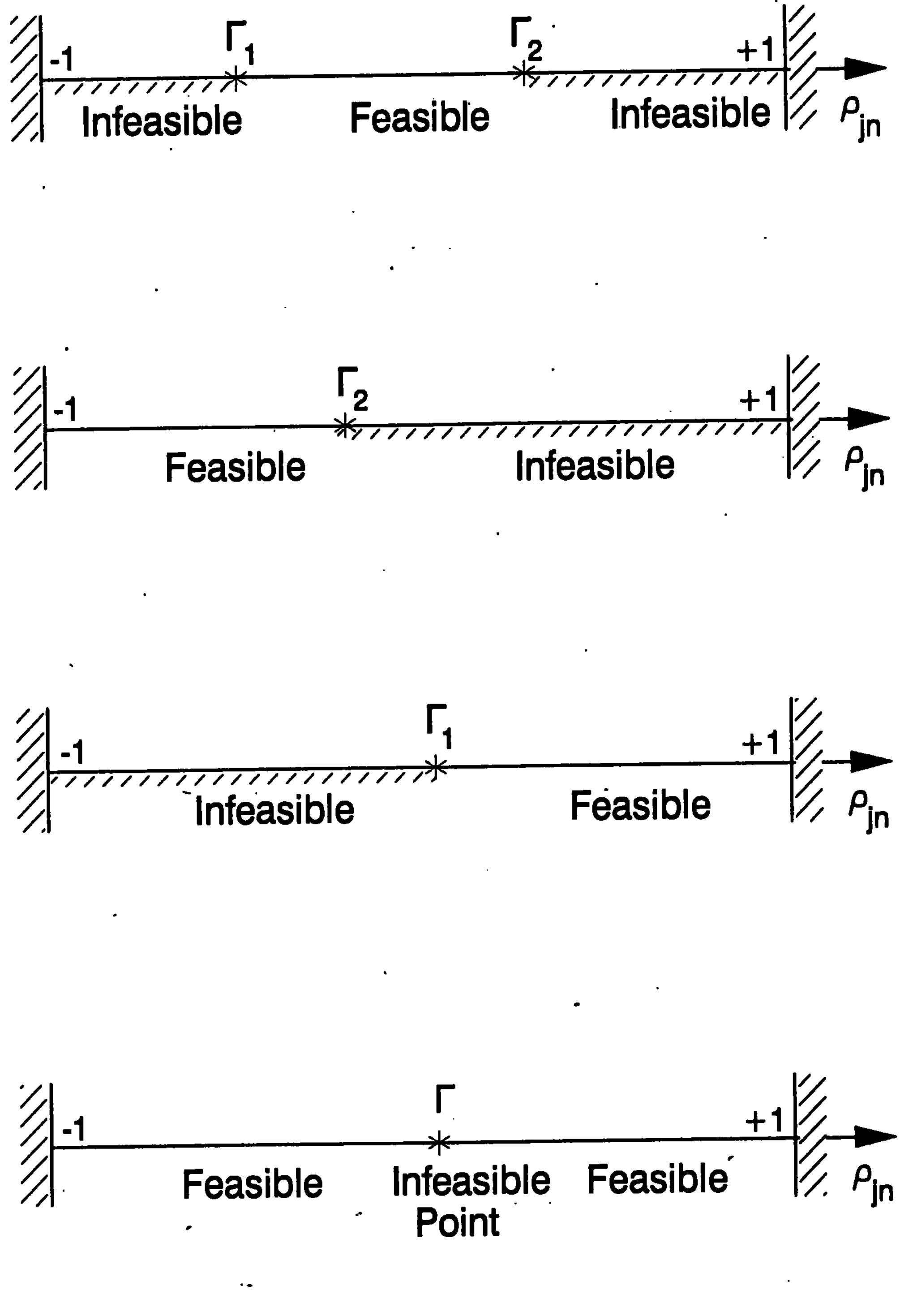

Fig. 1 Feasible Solution Regions for Γ

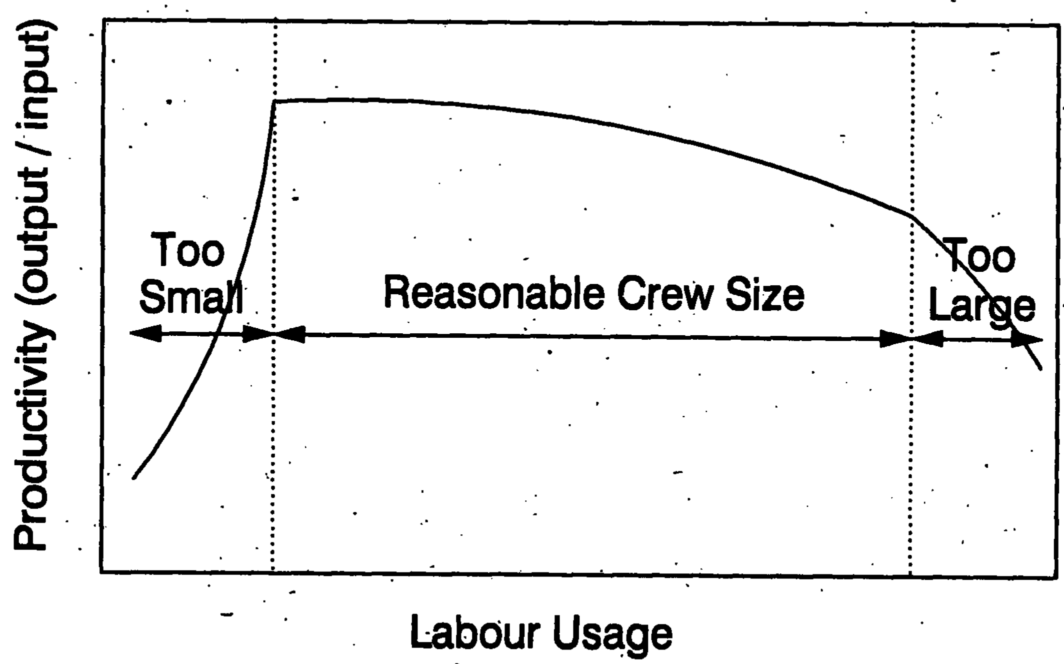

Fig. 2 Real Life Relationship Between *P & L*

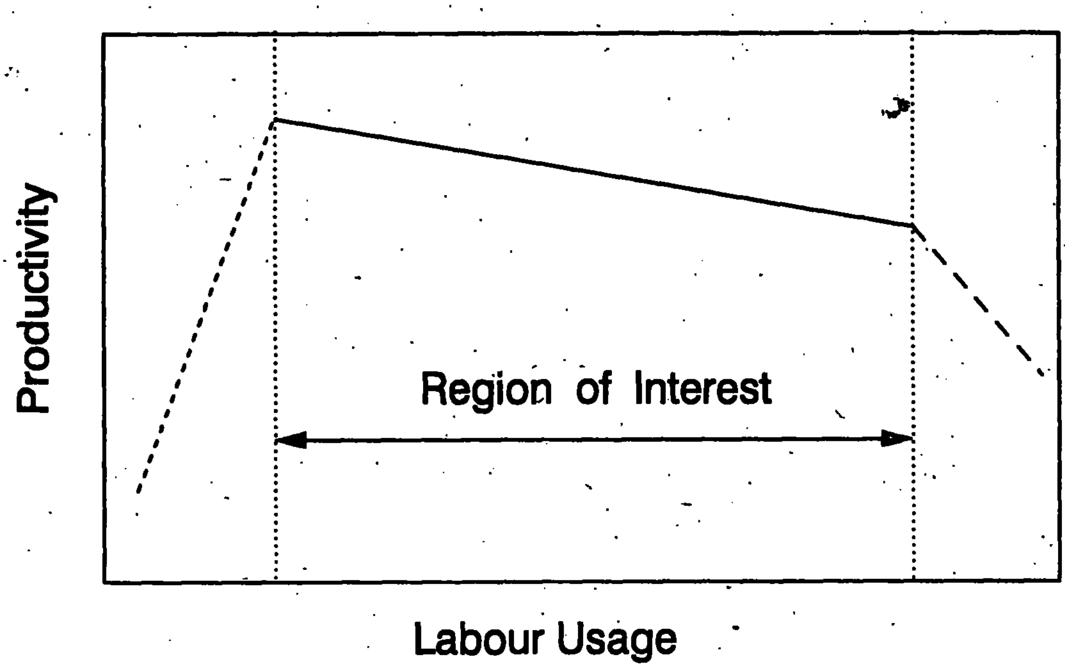

Fig. 3 Idealised Relationship Between *P & L*