5

The Algebra of Linear Regression and Partial Correlation

Our goal in this book is to study structural equation modeling in its full generality. However, much of our work will concentrate on "Linear Structural Relations Models," which might be described succinctly as multiple linear regression models with (possibly) latent independent and dependent variables. Just as structural equation modeling contains a linear regression model within its boundaries, so does factor analysis. Understanding the foundations of both factor analysis and structural equation modeling therefore requires an understanding of the key algebraic properties of linear regression. In this chapter, we begin by recalling the basic results in bivariate linear regression. We then proceed to multiple linear regression systems.

5.1 BIVARIATE LINEAR REGRESSION

In bivariate linear regression performed on a sample of N observations, we seek to examine the extent of the linear relationship between two observed variables, X and Y. One variable (usually the one labeled Y) is the *dependent* or *criterion* variable, the other (usually labeled X) is the *independent* or *predictor* variable. Each data point represents a pair of scores, x_i, y_i that may be plotted as a point in the plane. Such a plot, called a *scatterplot*, is shown in Figure 5.1. Then, a straight line is fitted to the data.

It would be a rare event, indeed, if all the points fell on a straight line. However, if Y and X have an approximate linear relationship, then a straight line, properly placed, should fall close to many of the points.

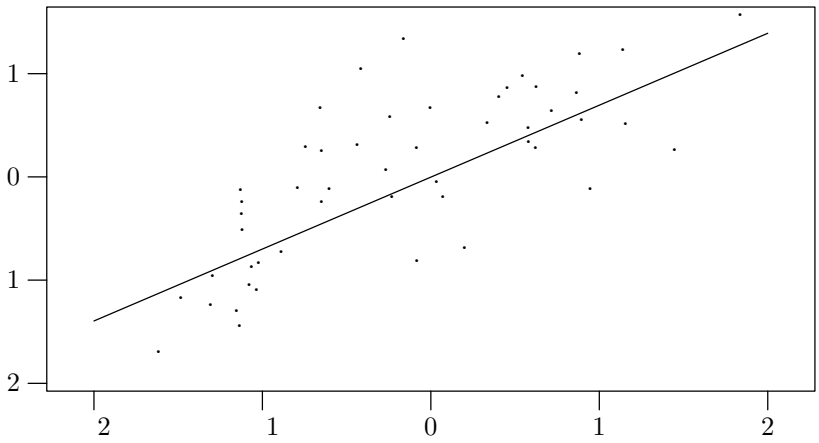
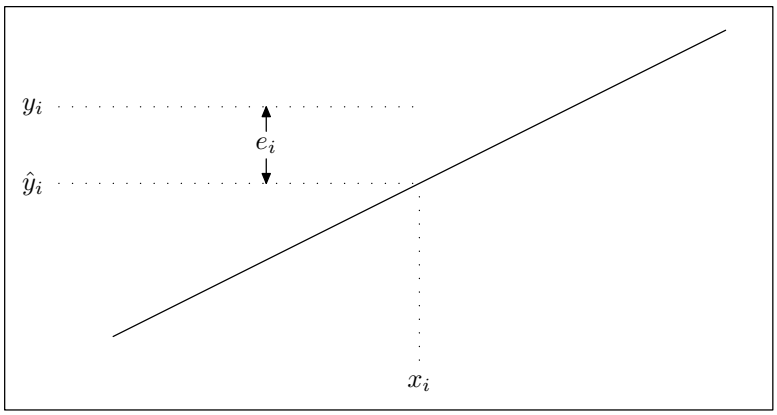


Fig. 5.1 A Scatterplot and Linear Regression Line

How does one decide when the line is positioned optimally? Generally, the *least squares* criterion is employed. Recall the basic notation of linear regression. The ith data point has coordinates (x_i, y_i) in the plane. Any "regression line" is fully specified by its slope b and Y-intercept c, i.e., it fits the equation $Y = bX + c$.

Figure 5.2 shows a single point in relation to a line in the plane. Suppose one were trying to use the regression line to "predict" (or guess) the Y value for this particular point from its X value, simply using the regression line. To do this, one would evaluate the straight line function at the X value, by going up to the line at x_i. Typical notation is to call the predicted Y value $\hat{y}_i$. It

Fig. 5.2 Linear Regression Notation



follows immediately that

$$\hat{y}_i = bx_i + c \tag{5.1}$$

and

$$y_i = \hat{y}_i + e_i \tag{5.2}$$

where e_i is defined tautologically as simply the (signed) distance in the up-down direction between the point and the line, i.e.,

$$e_i = y_i - \hat{y}_i \tag{5.3}$$

The *least squares criterion* chooses the best-fitting straight line by minimizing the sum of squared errors, i.e., $\sum_{i=1}^{N} e_i^2$. Since, for any fixed set of data points, the sum of squared errors is a function of the placement of the straight line, it may be viewed as a function in two unknowns, b and c. Minimizing a function of two unknowns is a straightforward exercise in differential calculus. The result is well known, i.e., for the best-fitting straight line, the slope b and Y-intercept c are

$$b = r_{yx}\frac{S_y}{S_x} \tag{5.4}$$

and

$$c = \overline{y}_\bullet - b\overline{x}_\bullet \tag{5.5}$$

Note that, if Y and X are expressed in Z-score form, then $b = r_{yx}$ and $c = 0$. Also note that, by substituting $S_{yx}S_yS_x$ for r_{yx} in Equation 5.4, one obtains an alternative expression for b, i.e.,

$$b = \frac{S_{yx}}{S_x^2} \tag{5.6}$$

If Y and X are expressed in deviation score form, the result of Equation 5.6 still holds, while $c = 0$.

Using linear transformation rules, one may derive expressions for the variance of the predicted ($\hat{y}_i$) scores, the error (e_i) scores, and the covariance between them. For example consider the variance of the predicted scores. Remember that adding a constant has no effect on a variance, and multiplying by a constant multiplies the variance by the square of the multiplier. So, since $\hat{y}_i = bx_i + c$, it follows immediately that

$$\begin{aligned} s_{\hat{y}}^2 &= b^2 S_x^2 \\ &= (r_{yx}S_y/S_x)^2 S_x^2 \\ &= r_{yx}^2 S_y^2 \end{aligned} \tag{5.7}$$

The covariance between the criterion scores (y_i) and predicted scores ($\hat{y}_i$) is obtained by the heuristic rule. Begin by re-expressing $\hat{y}_i$ as $bx_i + c$, then recall that additive constant c cannot affect a covariance. So the covariance between y_i and $\hat{y}_i$ is the same as the covariance between y_i and bx_i. Using the heuristic

approach, we find that $S_{y,\hat{y}} = S_{y,bx} = bS_{yx}$ Recalling that $S_{yx} = r_{yx}S_yS_x$, and $b = r_{yx}S_y/S_x$, one immediately arrives at

$$S_{y,\hat{y}} = r_{yx}^2 S_y^2 = S_{\hat{y}}^2 \tag{5.8}$$

Calculation of the covariance between the predicted scores and error scores proceeds in much the same way. Re-express e_i as $y_i - \hat{y}_i$, then use the heuristic rule. One obtains

$$\begin{aligned} S_{\hat{y},e} &= S_{\hat{y},y-\hat{y}} \\ &= S_{\hat{y},y} - S_{\hat{y}}^2 \\ &= S_{\hat{y}}^2 - S_{\hat{y}}^2 \quad \text{(from Equation 5.8)} \\ &= 0 \end{aligned} \tag{5.9}$$

Predicted and error scores always have exactly zero covariance, and zero correlation, in linear regression.

Using a similar approach, we may prove that

$$S_y^2 = S_{\hat{y}}^2 + S_e^2 \tag{5.10}$$

5.2 MULTIVARIATE LINEAR REGRESSION

Multiple linear regression with a single criterion variable is a straightforward generalization of linear regression. To make the notation simpler, assume that the criterion variable Y and the p predictor variables X_j, $j = 1, \ldots, p$ are in deviation score form.

Let $\mathbf{y}$ be an $N \times 1$ vector of criterion scores, and $\mathbf{X}$ be the $N \times p$ matrix with the predictor variables in columns. Then the multiple regression prediction equation is

$$\begin{aligned} \mathbf{y} &= \hat{\mathbf{y}} + \mathbf{e} \\ &= \mathbf{X}\mathbf{b} + \mathbf{e} \end{aligned} \tag{5.11}$$

The least squares criterion remains essentially as before, i.e., minimize $\mathbf{e}'\mathbf{e}$ under choice of $\mathbf{b}$. The unique solution is

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{5.12}$$

The above notation generalizes immediately to situations where more than one criterion is being predicted simultaneously. Specifically, let $N \times q$ matrix $\mathbf{Y}$ contain q criterion variables, and let $\mathbf{B}$ be a $p \times q$ matrix of regression weights. The least squares criterion is satisfied when the sum of squared errors across all variables (i.e. $\mathrm{Tr}\left(\mathbf{E}'\mathbf{E}\right)$) is minimized. The unique solution is the obvious generalization of Equation 5.12, i.e.,

$$\mathbf{B} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \tag{5.13}$$

We will now prove some multivariate generalizations of the properties we developed earlier for bivariate linear regression systems. First, we prove that $\hat{\mathbf{Y}} = \mathbf{XB}$ and $\mathbf{E} = \mathbf{Y} - \mathbf{XB}$ are uncorrelated. To do this, we examine the covariance matrix between them, and prove that it is a null matrix. Recall from page 73 that, when scores in $\mathbf{Y}$ and $\mathbf{X}$ are in deviation score form, that $\mathbf{S_{yx}} = 1/(N-1)\mathbf{Y}'\mathbf{X}$. Hence, (moving the $N-1$ to the left of the formula for simplicity),

$$\begin{aligned}
(N-1)\mathbf{S}_{\hat{\mathbf{y}},\mathbf{e}} &= \hat{\mathbf{Y}}'\mathbf{E} \\
&= (\mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\
&= \mathbf{B}'\mathbf{X}'(\mathbf{Y} - \mathbf{XB}) \\
&= \mathbf{B}'\mathbf{X}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{XB} \\
&= \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
&= 0 \qquad (5.14)
\end{aligned}$$

The preceding result makes it easy to show that the variance-covariance matrix of $\mathbf{Y}$ is the sum of the variance-covariance matrices for $\hat{\mathbf{Y}}$ and $\mathbf{E}$. Specifically,

$$\begin{aligned}
\mathbf{S_{yy}} &= 1/(N-1)\mathbf{Y}'\mathbf{Y} \\
&= 1/(N-1)\left(\hat{\mathbf{Y}} + \mathbf{E}\right)'\left(\hat{\mathbf{Y}} + \mathbf{E}\right) \\
&= 1/(N-1)\left(\hat{\mathbf{Y}}' + \mathbf{E}'\right)\left(\hat{\mathbf{Y}} + \mathbf{E}\right) \\
&= 1/(N-1)\hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\hat{\mathbf{Y}} + \hat{\mathbf{Y}}'\mathbf{E} + \mathbf{E}'\mathbf{E} \\
&= 1/(N-1)\hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{0} + \mathbf{0} + \mathbf{E}'\mathbf{E} \\
&= 1/(N-1)\hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{E}'\mathbf{E} \\
&= \mathbf{S}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} + \mathbf{S_{ee}} \qquad (5.15)
\end{aligned}$$

Notice also that

$$\mathbf{S_{ee}} = \mathbf{S_{yy}} - \mathbf{B}'\mathbf{S_{xx}}\mathbf{B} \qquad (5.16)$$

Similar relationships hold when systems of random variables are related in a least-squares multiple regression setup. Specifically, suppose there are p criterion variables in the random vector $\boldsymbol{\eta}$, and q predictor variables in the random vector $\boldsymbol{\xi}$. The prediction equation is

$$\begin{aligned}
\boldsymbol{\eta} &= \mathbf{B}'\boldsymbol{\xi} + \boldsymbol{\epsilon} \qquad (5.17) \\
&= \hat{\boldsymbol{\eta}} + \boldsymbol{\epsilon} \qquad (5.18)
\end{aligned}$$

In the population, the least-squares solution minimizes the average squared error, i.e., $\mathrm{Tr}\,(\mathcal{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'))$. The solution for $\mathbf{B}$ is

$$\mathbf{B} = \boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\eta}} \tag{5.19}$$

The covariance matrix between predicted and error variables is null, just as in the sample case. The proof is structurally similar to its sample counterpart, but we include it here to demonstrate several frequently used techniques in the matrix algebra of expected values.

$$\begin{aligned}
\boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}\boldsymbol{\epsilon}} &= \mathcal{E}\,(\hat{\boldsymbol{\eta}}\boldsymbol{\epsilon}') \\
&= \mathcal{E}\left(\mathbf{B}'\boldsymbol{\xi}(\boldsymbol{\eta} - \mathbf{B}'\boldsymbol{\xi})'\right) \\
&= \mathcal{E}\left(\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\xi}\boldsymbol{\eta}' - \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\right) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\mathcal{E}(\boldsymbol{\xi}\boldsymbol{\eta}') - \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\mathcal{E}(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\eta}} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\eta}} - \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\xi}} \\
&= \mathbf{0}
\end{aligned} \tag{5.20}$$

We also find that

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\eta}} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}} \tag{5.21}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\eta}} - \mathbf{B}'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\mathbf{B} \tag{5.22}$$

Now consider an individual random variable η_i in $\boldsymbol{\eta}$. The correlation between η_i and its respective $\hat{\eta}_i$ is called "the multiple correlation of η_i with $\boldsymbol{\xi}$." Now, suppose that the variables in $\boldsymbol{\xi}$ were uncorrelated, and that they and the variables in $\boldsymbol{\eta}$ have unit variances, so that $\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}} = \mathbf{I}$, an identity matrix, and, as a consequence, $\mathbf{B} = \boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\eta}}$. Then the correlation between a particular

η_i and its respective $\hat{\eta}_i$ is

$$\begin{aligned} r_{\eta_i,\hat{\eta}_i} &= \frac{\sigma_{\eta_i\hat{\eta}_i}}{\sqrt{\sigma^2_{\eta_i}\sigma^2_{\hat{\eta}_i}}} \\ &= \frac{\mathcal{E}\left(\eta_i(\mathbf{b}_i'\boldsymbol{\xi})'\right)}{\sqrt{(1)(\mathbf{b}_i'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\mathbf{b}_i)}} \\ &= \frac{\mathcal{E}\left(\eta_i\boldsymbol{\xi}'\mathbf{b}_i\right)}{\sqrt{(\mathbf{b}_i'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\mathbf{b}_i)}} \\ &= \frac{\mathcal{E}\left(\eta_i\boldsymbol{\xi}'\right)\mathbf{b}_i}{\sqrt{(\mathbf{b}_i'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\mathbf{b}_i)}} \\ &= \frac{\sigma_{\eta_i\boldsymbol{\xi}}\mathbf{b}_i}{\sqrt{(\mathbf{b}_i'\mathbf{b}_i)}} \\ &= \frac{\mathbf{b}_i'\mathbf{b}_i}{\sqrt{(\mathbf{b}_i'\mathbf{b}_i)}} \end{aligned} \tag{5.23}$$

It follows immediately that, when the predictor variables in $\boldsymbol{\xi}$ are orthogonal with unit variance, squared multiple correlations may be obtained directly as a sum of squared, standardized regression weights.

In subsequent chapters, we will be concerned with two linear regression prediction systems known (loosely) as "factor analysis models," but referred to more precisely as "common factor analysis" and "principal component analysis." In each system, we will be attempting to reproduce an observed (or "manifest") set of p random variables in as (least squares) linear functions of a smaller set of m hypothetical (or "latent") random variables.

5.3 PARTIAL CORRELATION

In many situations, the correlation between two variables may be substantially different from zero without implying any causal connection between them. A classic example is the high positive correlation between number of fire engines sent to a fire and the damage done by the fire. Clearly, sending fire engines to a fire does not usually cause damage, and it is equally clear that one would be ill-advised to recommend reducing the number of trucks sent to a fire as a means of reducing damage. In situations like the example above, one looks for (indeed often hypothesizes on theoretical grounds) a "third variable" which is causally connected with the first two variables, and "explains" the correlation between them. In the above example, such a third variable might be "size of fire." One would expect that, if size of fire were held constant, there would be, if anything, a negative correlation between damage done by a fire and the number of fire engines sent to the fire. One way of statistically

holding the third variable "constant" is through partial correlation analysis. In this analysis, we "partial out" the third variable from the first two by linear regression, leaving two linear regression error, or *residual* variables. We then compute the "partial correlation" between the first two variables as the correlation between the two regression residuals. A basic notion connected with partial correlation analysis is that, if, by partialling out one or more variables, you cause the partial correlations among some (other) variables to go to zero, then you have "explained" the correlations among the (latter) variables as being "due to" the variables which were partialled out. If, in terms of Equation 5.18 above, we "explain" the correlations in the variables in $\boldsymbol{\eta}$ by the variables in $\boldsymbol{\xi}$, then $\boldsymbol{\epsilon}$ should have a correlation (and covariance) matrix which is diagonal, i.e., the variables in $\boldsymbol{\eta}$ should be uncorrelated once we "partial out" the variables in $\boldsymbol{\xi}$ by linear regression. Recalling Equation 5.22 we see that this implies that $\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\eta}} - \mathbf{B}'\boldsymbol{\Sigma}_{\boldsymbol{\xi}\boldsymbol{\xi}}\mathbf{B}$ is a diagonal matrix.

This seemingly simple result has some rather surprisingly powerful ramifications, once one drops certain restrictive mental sets. In the next chapter, we see how, at the turn of the 20th century, it led Charles Spearman to a revolutionary model for human intelligence, and an important new statistical technique for testing the model with data. What was surprising about the model was that it could be tested, even though the predictor variables (in $\boldsymbol{\xi}$) are never directly observed!

Problems

5.1 Suppose you have a system of variables $\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$, in which $\mathbf{x}$ is a set of standardized, orthogonal predictors (i.e., their covariance matrix is $\mathcal{E}(\mathbf{x}\mathbf{x}') = \mathbf{I}$), $\mathbf{F}$ is a set of least squares regression weights for predicting $\mathbf{y}$ from $\mathbf{x}$, *and* the predictors in $\mathbf{x}$ are known to be orthogonal to the residuals in $\mathbf{e}$. Furthermore, assume that the residuals have a *diagonal variance-covariance matrix* $\mathbf{D}$.

5.1.1. Find a simple expression for the covariance matrix between y and x.

5.1.2. Can you imagine two real-world situations where a system like the one described in the above problem would occur? Remember that the fundamental features of the system are that the "signal" ($\mathbf{x}$) is orthogonal to the noise ($\mathbf{e}$), and that the signal explains the correlations in the observed variables (in the partial correlation sense).

5.2 Find an expression for the variance-covariance matrix of $\mathbf{y}$ in terms of $\mathbf{F}$ and $\mathbf{D}$. (*Hint.* Write $\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e}$ and compute $\boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}} = \mathcal{E}(\mathbf{y}\mathbf{y}')$.)