$$P = g(\mathbf{m}) + (\mathbf{X} - \mathbf{m})' \partial_{\mathbf{x}} g|_{\mathbf{x}=\mathbf{m}} + \frac{1}{2}(\mathbf{X} - \mathbf{m})' \partial^2_{\mathbf{xx}'} g|_{\mathbf{x}=\mathbf{m}} (\mathbf{X} - \mathbf{m}) + \cdots, \tag{3.108}$$

where $\mathbf{m}$ is a significative value of the invariants. One standard choice is zero:

$$\mathbf{m} \equiv \mathbf{0}. \tag{3.109}$$

Another standard choice is the expected value:

$$\mathbf{m} \equiv \mathrm{E}\{\mathbf{X}\}. \tag{3.110}$$

If the approximation in (3.108) is performed up to the first order, the market prices at the horizon are a linear function of the invariants. If the approximation is carried on up to the second order, the market prices are quadratic functions of the invariants. In either case, the distribution of the market prices becomes a tractable expression of the distribution of the invariants.

Depending on its end users, the approximation (3.108) is known under different names.

In the derivatives world the expansion up to order zero is called the *theta approximation*. The expansion up to order one is called the *delta-vega approximation*. The *delta* is the first derivative (mathematical operation) of the investment-horizon pricing function of the derivative (financial contract) with respect to the underlying, whereas the *vega* is the first derivative (mathematical operation) of the investment-horizon pricing function of the derivative (financial contract) with respect to the implied volatility. The expansion up to order two is called the *gamma approximation*. The *gamma* is the second derivative (mathematical operation) of the investment-horizon pricing function of the derivative (financial contract) with respect to the underlying.

In the fixed-income world the expansion up to order zero in (3.108) is known as the *roll-down* or *slide approximation*. The expansion up to order one is known as the *PVBP* or *duration approximation*. The expansion up to order two is known as the *convexity approximation*, see Section 3.5 for a thorough case-study.

We stress again that the accuracy of (3.108) is jeopardized by the hidden threat of *estimation risk*, which we discuss in the third part of the book.

3.4 Dimension reduction

According to (3.79), the prices at the investment horizon of the securities in our market are a function of the randomness in the market:

$$\mathbf{P}_{T+\tau} = \mathbf{g}(\mathbf{X}_{T+\tau,\tau}), \tag{3.111}$$

where $\mathbf{X}_{t,\tau}$ denotes the generic set of market invariants relative to the interval τ that becomes known at time t.

In a generic market of a large number of securities, the following two phenomena typically occur.

In the first place the actual dimension of the market is less than the number of securities. This is due to the joint presence in the market of derivatives and underlying securities. Such phenomena can be analyzed in terms of the copula of the market and the related dependence summary statistics, as discussed in Section 2.5.

For example, consider a market of two products: a stock which trades at the generic time t at the price S_t and a call option on that stock with strike $\widetilde{S}$ that trades at the price C_t. If the investment horizon coincides with the expiry of the option, the market is one-dimensional:

$$\mathbf{P}_{T+\tau} = \begin{pmatrix} S_{T+\tau} \\ \max\left(S_{T+\tau} - \widetilde{S}, 0\right) \end{pmatrix}, \tag{3.112}$$

see Figure 2.5. From Table 2.115, the Schweizer and Wolff measure of dependence between these two securities is one, i.e. the maximum possible value.

In the second place, the actual dimension of the randomness in the market, i.e. the actual dimension of the N-dimensional vector of investment-horizon invariants $\mathbf{X}$, is less than N. This is the subject of the remainder of this section.

We aim at expressing the vector of invariants $\mathbf{X}$ as a function of two sets of variables: a vector $\mathbf{F}$ of a few *common factors* that are responsible for most of the randomness in the market; and a residual vector $\mathbf{U}$ of *perturbations* that have a marginal effect:

$$\mathbf{X}_{t,\tau} \equiv \mathbf{h}\left(\mathbf{F}_{t,\tau}\right) + \mathbf{U}_{t,\tau}. \tag{3.113}$$

In this expression the vector of factors $\mathbf{F}$ should have a much lower dimension than the market invariants:

$$K \equiv \dim\left(\mathbf{F}_{t,\tau}\right) \ll N \equiv \dim\left(\mathbf{X}_{t,\tau}\right). \tag{3.114}$$

We remark that, since $\mathbf{X}_{t,\tau}$ represents the market invariants, i.e. it is a vector of independent and identically distributed random variables that become known at time t, both factors $\mathbf{F}_{t,\tau}$ and perturbations $\mathbf{U}_{t,\tau}$ must also be market invariants. In the sequel we drop the generic time t and the generic interval τ from the notation.

Intuitively, the factors should affect all the invariants and be responsible for most of the randomness in the market. In other words the invariants recovered through the factors should be very close to the original market invariants:

$$\widetilde{\mathbf{X}} \equiv \mathbf{h}\left(\mathbf{F}\right) \approx \mathbf{X}. \tag{3.115}$$

To measure the goodness of this approximation we use the *generalized r-square*, which we define as follows:

$$R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}\right\} \equiv 1 - \frac{\mathrm{E}\left\{\left(\mathbf{X} - \widetilde{\mathbf{X}}\right)' \left(\mathbf{X} - \widetilde{\mathbf{X}}\right)\right\}}{\mathrm{tr}\left\{\mathrm{Cov}\left\{\mathbf{X}\right\}\right\}}. \tag{3.116}$$

The term in the numerator is a measure of the amount of randomness in the residual, which is zero if and only if the approximation (3.115) is exact. The term in the denominator is a measure of the amount of randomness in the original invariants, as it is proportional to the average of the variances of all the invariants. The factor model (3.113) is viable if the generalized r-square approaches one. An r-square close to zero or even negative indicates that the factor model performs poorly.

The generic factor model (3.113) is too broad. In the sequel we will restrict our models to linear functions. In other words, we express the invariants in the following form:

$$\mathbf{X} \equiv \mathbf{B}\mathbf{F} + \mathbf{U}. \tag{3.117}$$

The K columns of the $N \times K$ matrix $\mathbf{B}$ are called the *factor loadings*: they transfer the effect of each of the K factors in $\mathbf{F}$ to the N invariants in $\mathbf{X}$. Notice that (3.117) represents a first-order Taylor approximations of the general formula (3.113), if we include a constant among the factors.

Ideally, common factors and perturbations should be independent variables. For practical purposes this requirement is too restrictive, therefore we only impose that common factors and perturbation be uncorrelated:

$$\mathrm{Cor}\left\{\mathbf{F}, \mathbf{U}\right\} = \mathbf{0}_{K \times N}, \tag{3.118}$$

which is a weaker assumption, see (2.136). The two assumptions (3.117) and (3.118) encompass the vast majority of the factor models considered in the financial literature.

Factor models for the market invariants can be obtained in two ways: either the factors are measurable market invariants, in which case we obtain an explicit factor model, or they are synthetic variables defined in terms of the original market invariants, in which case we obtain a hidden factor model. In either case, the perturbations are defined as the residual term.

3.4.1 Explicit factors

Here we assume that the factors $\mathbf{F}$ in the linear factor model (3.117) are explicit market variables. In other words, for any choice of the $N \times K$ matrix $\mathbf{B}$ of the factor loadings we obtain a linear model that defines the residuals as follows:

$$\mathbf{X} \equiv \mathbf{B}\mathbf{F} + \mathbf{U}. \tag{3.119}$$

The *regression factor loadings* correspond to the best choice of the coefficients $\mathbf{B}$ in terms of the generalized r-square criterion (3.116). By definition the regression factor loadings solve:

$$\mathbf{B}_r \equiv \underset{\mathbf{B}}{\operatorname{argmax}}\, R^2\left\{\mathbf{X}, \mathbf{B}\mathbf{F}\right\}, \tag{3.120}$$

where "r" stands for "regression".

As we show in Appendix www.3.4, the regression factor loadings read:

$$\mathbf{B}_r \equiv \mathrm{E}\left\{\mathbf{X}\mathbf{F}'\right\}\mathrm{E}\left\{\mathbf{F}\mathbf{F}'\right\}^{-1}. \tag{3.121}$$

The regression factor loadings in turn yield the recovered invariants $\widetilde{\mathbf{X}}_r \equiv \mathbf{B}_r\mathbf{F}$ and the perturbations, i.e. the residuals $\mathbf{U}_r \equiv \mathbf{X} - \widetilde{\mathbf{X}}_r$. Unfortunately, the perturbations do not display zero correlation with the explicit factors unless the factors have zero expected value:

$$\mathrm{E}\left\{\mathbf{F}\right\} = \mathbf{0} \Rightarrow \mathrm{Cor}\left\{\mathbf{F}, \mathbf{U}\right\} = \mathbf{0}_{K\times N}. \tag{3.122}$$

For example, consider an invariant and a factor that are jointly normally distributed:

$$\begin{pmatrix} X \\ F \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_X \\ \mu_F \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_F \\ \rho\sigma_X\sigma_F & \sigma_F^2 \end{pmatrix}\right). \tag{3.123}$$

In this case the regression factor loading reads:

$$b_r = \frac{\mu_X\mu_F + \rho\sigma_X\sigma_F}{\mu_F^2 + \sigma_F^2}. \tag{3.124}$$

From the more general formulas of Appendix www.3.4 we obtain:

$$\mathrm{Cov}\left\{U, F\right\} = \rho\sigma_X\sigma_F\left(1 - \frac{1}{1+\mu_F^2/\sigma_F^2}\right) - \frac{\mu_F\mu_X}{1+\mu_F^2/\sigma_F^2}, \tag{3.125}$$

which is null if $\mu_F \equiv 0$.

Nevertheless, we can always include a constant among the factors:

$$\mathbf{F} \mapsto \begin{pmatrix} 1 \\ \mathbf{F} \end{pmatrix}. \tag{3.126}$$

We show in Appendix www.3.4 that in this case the regression coefficients (3.121) yield the following recovered invariants:

$$\widetilde{\mathbf{X}}_r \equiv \mathrm{E}\left\{\mathbf{X}\right\} + \mathrm{Cov}\left\{\mathbf{X}, \mathbf{F}\right\}\mathrm{Cov}\left\{\mathbf{F}\right\}^{-1}\left(\mathbf{F} - \mathrm{E}\left\{\mathbf{F}\right\}\right), \tag{3.127}$$

see Figure 3.12. The perturbations, which are defined as the residuals $\mathbf{U}_r \equiv \mathbf{X} - \widetilde{\mathbf{X}}_r$, have zero expected value and display zero correlation with the factors:

$$\mathrm{E}\left\{\mathbf{U}_r\right\} = \mathbf{0}, \quad \mathrm{Cor}\left\{\mathbf{F}, \mathbf{U}\right\} = \mathbf{0}_{K\times N}. \tag{3.128}$$

Furthermore, the covariance of the residual reads:

$$\mathrm{Cov}\left\{\mathbf{U}_r\right\} = \mathrm{Cov}\left\{\mathbf{X}\right\} - \mathrm{Cov}\left\{\mathbf{X}, \mathbf{F}\right\}\mathrm{Cov}\left\{\mathbf{F}\right\}^{-1}\mathrm{Cov}\left\{\mathbf{F}, \mathbf{X}\right\}. \tag{3.129}$$

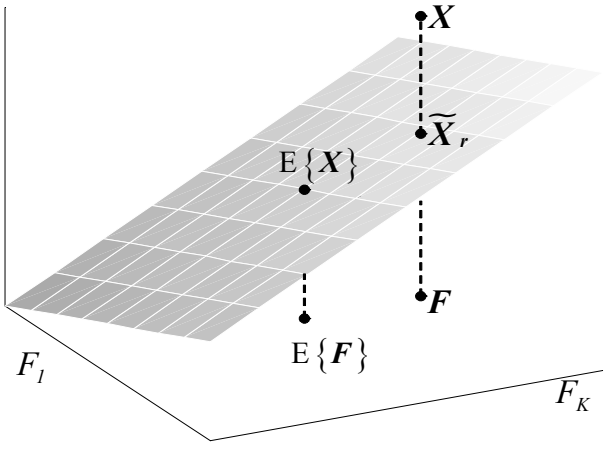

Fig. 3.12. Explicit factor dimension reduction: regression

Notice the similarities between the recovered invariants (3.127) and the expected value of the conditional normal distribution (2.165) on the one hand, and the covariance of the residuals (3.129) and the covariance of the conditional normal distribution (2.166) on the other hand.

If in our example (3.123) we add a constant we obtain from (3.127) the following recovered invariant:

$$\widetilde{X}_r \equiv \mu_X + \rho\frac{\sigma_X}{\sigma_F}(F - \mu_F). \tag{3.130}$$

This is the expected value of the conditional distribution of the invariant given the factor (2.174). Similarly, the variance of the residual reads:

$$\mathrm{Var}\left\{\widetilde{U}_r\right\} = \sigma_X^2\left(1 - \rho^2\right), \tag{3.131}$$

which is the variance of the conditional distribution of the invariant given the factor (2.175).

In order to evaluate the quality of an explicit factor model, it is better to reformulate our model in a scale-independent fashion.

First of all we normalize the market invariants by means of their z-scores $\mathbf{Z}_X$, which from (1.35) read component-wise:

$$Z_X^{(n)} \equiv \frac{X_n - \mathrm{E}\{X_n\}}{\sqrt{\mathrm{Cov}\{X_n, X_n\}}}. \tag{3.132}$$

The z-scores have zero expected value and unit standard-deviation: therefore they represent a scale- and location-independent version of the market invariants.

To normalize the factors, we consider the principal component decomposition (2.76) of their covariance matrix:

$$\mathrm{Cov}\{\mathbf{F}\} \equiv \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'. \tag{3.133}$$

In this expression $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues sorted in decreasing order:

$$\boldsymbol{\Lambda} \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_K); \tag{3.134}$$

and $\mathbf{E}$ is the juxtaposition of the respective eigenvectors:

$$\mathbf{E} \equiv \left(\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(K)}\right). \tag{3.135}$$

This matrix satisfies $\mathbf{E}\mathbf{E}' = \mathbf{I}_K$ and thus it represents a rotation, see Figure A.4. In terms of the principal component decomposition we can normalize the factors as follows:

$$\mathbf{Z}_F \equiv \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{E}'(\mathbf{F} - \mathrm{E}\{\mathbf{F}\}). \tag{3.136}$$

These are the z-scores of the factors, rotated in a way that decorrelates them:

$$\mathrm{Cov}\{\mathbf{Z}_F\} = \mathbf{I}_K, \tag{3.137}$$

see the proof in Appendix www.3.4.

In terms of the normalized variables (3.132) and (3.136), the recovered invariants (3.127) read:

$$\widetilde{\mathbf{Z}}_X = \mathbf{C}_{XF}\mathbf{Z}_F, \tag{3.138}$$

where the matrix $\mathbf{C}_{XF}$ is the correlation between the market invariants and the (rotated) explicit factors:

$$\mathbf{C}_{XF} \equiv \mathrm{Cor}\{\mathbf{X}, \mathbf{E}'\mathbf{F}\}. \tag{3.139}$$

The correlation $\mathbf{C}_{XF}$ in (3.138) is responsible for transferring the randomness of the factors into the recovered invariants. Indeed, we show in Appendix www.3.4 that the generalized r-square (3.116) of the explicit factor model can be expressed as an average correlation:

$$R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}_r\right\} = \frac{\mathrm{tr}(\mathbf{C}_{XF}\mathbf{C}'_{XF})}{N}. \tag{3.140}$$

Therefore the factors should be chosen as correlated as possible to the market invariants, in order to increase their explanatory power.

In our example (3.123), where there exists only one factor, (3.139) reads:

$$\mathbf{C}_{XF} \equiv \mathrm{Cor}\{X, F\} = \rho. \tag{3.141}$$

Therefore in our simple one-factor model the generalized r-square (3.140) is the square of the correlation between the factor and the invariant:

$$R^2 = \rho^2. \tag{3.142}$$

Indeed, from (3.131) when the factor and the invariant are highly correlated, the residual is minimal and thus the explanatory power of the factor model is maximal.

Adding factors trivially improves the quality of the result. Nevertheless, the number of factors should be kept at a minimum, in order not to defeat the purpose of dimension reduction.

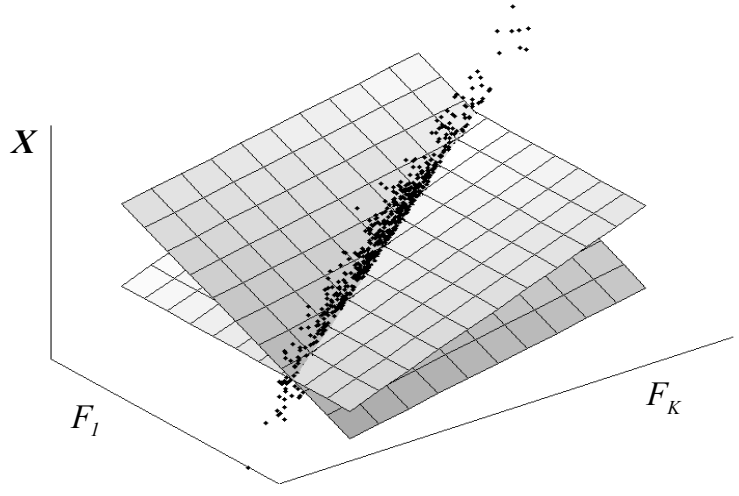

Fig. 3.13. Collinearity: the regression plane is not defined

Furthermore, the factors should be chosen as diversified as possible, in order to avoid the problem of *collinearity*. Indeed, when the K factors are not diversified they span a hyperplane of dimension less than K. This makes it impossible to identify the regression hyperplane, see Figure 3.13.

Several criteria have been developed in the statistical and financial literature to select the most suitable among a pool of potential explicit factors, such as the *Akaike information criterion* and the *Bayesian information criterion*. We refer the reader to references such as Parzen, Tanabe, and Kitagawa

(1998), see also Connor and Korajczyk (1993) for financial applications. To implement the selection in practice once a suitable criterion has been determined see Section 3.4.5.

3.4.2 Hidden factors

In a linear model with hidden factors we assume that the factors are not explicit market variables. Instead, they are functions of the original invariants that summarize as much information about the invariants as possible. Including a constant among the hidden factors, (3.117) reads:

$$\mathbf{X} \equiv \mathbf{q} + \mathbf{B}\mathbf{F}(\mathbf{X}) + \mathbf{U}. \tag{3.143}$$

For any choice of the constant $\mathbf{q}$ and of the factor loading matrix $\mathbf{B}$ and for any choice $\mathbf{F}(\cdot)$ of the functional form that summarizes the invariants $\mathbf{X}$ into the synthetic factors, we obtain a model that defines the residuals $\mathbf{U}$.

According to the r-square criterion (3.116) the best, yet trivial, joint choice of constant, factor loadings and functional form for the hidden factors is represented by $\mathbf{q} \equiv \mathbf{0}$, $\mathbf{B} \equiv \mathbf{I}$ and $\mathbf{F}(\mathbf{X}) \equiv \mathbf{X}$ respectively. In this case the residuals are null and thus the generalized r-square is one. Nevertheless, no dimension reduction takes place, i.e. (3.114) is not satisfied, since the number of factors is equal to the number of invariants.

Once we impose the condition that the number of hidden factors be less than the number of invariants, the "best" linear model depends on the possible functional form that we consider for the factors.

Here we present two choices for the above functional form, which give rise to two approaches to hidden factor dimension reduction: principal component analysis and idiosyncractic factors.

Principal component analysis

Principal component analysis (*PCA*) provides the best dimension reduction under the assumption that the hidden factors in (3.143) be affine transformations of the invariants:

$$\mathbf{F}_p \equiv \mathbf{d}_p + \mathbf{A}_p' \mathbf{X}, \tag{3.144}$$

where $\mathbf{d}$ is a K-dimensional vector, $\mathbf{A}$ is an $N \times K$ matrix and "p" stands for "PCA". Notice that this is a first-order Taylor expansion of the more general functional form $\mathbf{F}(\mathbf{X})$ that appears in (3.143).

Under the above assumption, from (3.143) the optimally recovered invariants must be an affine transformation of the original invariants:

$$\widetilde{\mathbf{X}}_p \equiv \mathbf{m}_p + \mathbf{B}_p \mathbf{A}_p' \mathbf{X}, \tag{3.145}$$

where

$$\mathbf{m}_p \equiv \mathbf{q} + \mathbf{B}_p \mathbf{d}_p. \tag{3.146}$$

Therefore, the PCA solution is represented by the following set of factor loadings and coefficients:

$$(\mathbf{B}_p, \mathbf{A}_p, \mathbf{m}_p) \equiv \underset{\mathbf{B},\mathbf{A},\mathbf{m}}{\operatorname{argmax}} R^2 \left\{\mathbf{X}, \mathbf{m} + \mathbf{B}\mathbf{A}'\mathbf{X}\right\}. \tag{3.147}$$

From this solution we can identify the coefficients $\mathbf{q}$ and $\mathbf{d}$ by imposing for instance the following condition:

$$\mathrm{E}\{\mathbf{F}\} \equiv \mathbf{0}. \tag{3.148}$$

To present the solution to this problem, we consider the spectral decomposition of the covariance matrix (2.76), which we report here:

$$\mathrm{Cov}\{\mathbf{X}\} \equiv \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'. \tag{3.149}$$

In this expression $\boldsymbol{\Lambda}$ is the diagonal matrix of the decreasing, positive eigenvalues of the covariance:

$$\boldsymbol{\Lambda} \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_N); \tag{3.150}$$

and $\mathbf{E}$ is the juxtaposition of the respective eigenvectors:

$$\mathbf{E} \equiv \left(\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(N)}\right), \tag{3.151}$$

which satisfies $\mathbf{E}\mathbf{E}' = \mathbf{I}_N$. Also, we consider the location-dispersion ellipsoid (2.75) associated with the expected value and the covariance matrix, see Figure 3.14.

First, we present a heuristic argument under the assumption that we only require one factor, i.e. $K \equiv 1$. We guess that this factor reads:

$$F \equiv \left[\mathbf{e}^{(1)}\right]' \mathbf{X}. \tag{3.152}$$

Indeed, from (2.82)-(2.83) the one-dimensional variable F captures the most randomness contained in the invariants that is possible by means of a linear transformation. The variable F represents the orthogonal projection of the variable $\mathbf{X}$ onto the direction defined by the first eigenvector, i.e. the longest principal axis in the location-dispersion ellipsoid.

To recover the N-dimensional invariant $\mathbf{X}$ with an affine transformation of F we must proceed as follows: we choose a fixed vector, i.e. a direction in $\mathbb{R}^N$; we multiply this vector by F; and we add a constant vector $\mathbf{m}$, i.e. we "center" the newly defined recovered variable.

Since the direction that contains most of the randomness in $\mathbf{X}$ is the longest principal axis, we let the random variable F vary along that direction by multiplying it by the first eigenvector $\mathbf{e}^{(1)}$. From (3.152) this means that the recovered invariants become the following affine function of the original invariants:

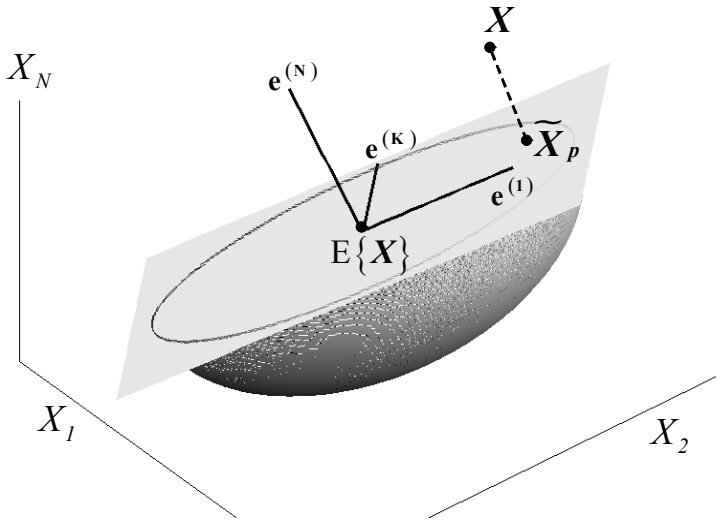

Fig. 3.14. Hidden factor dimension reduction: PCA

$$\widetilde{\mathbf{X}} \equiv \mathbf{m} + \mathbf{e}^{(1)} \left[\mathbf{e}^{(1)}\right]' \mathbf{X}. \tag{3.153}$$

To properly choose $\mathbf{m}$, i.e. to properly center the above recovered invariants, we impose that the expected values of both the original and the recovered invariants be the same. From this condition we immediately obtain:

$$\mathbf{m} \equiv \left(\mathbf{I}_N - \left[\mathbf{e}^{(1)}\right] \left[\mathbf{e}^{(1)}\right]'\right) \mathrm{E}\{\mathbf{X}\}, \tag{3.154}$$

where $\mathbf{I}_N$ is the identity matrix. Notice that with this choice of $\mathbf{m}$ the optimally recovered invariants (3.153) become the orthogonal projection of the original invariants along the direction of the longest principal axis of the location-dispersion ellipsoid. This is the line that contains the maximum possible randomness of the original invariants, i.e. the line that contains the maximum information about the original invariants.

Since (3.153)-(3.154) are in the form (3.145) we would argue that they provide the PCA dimension reduction (3.147) by means of one factor:

$$\{\mathbf{B}_p, \mathbf{A}_p, \mathbf{m}_p\} \equiv \left\{\mathbf{e}^{(1)}, \mathbf{e}^{(1)}, \left(\mathbf{I}_N - \left[\mathbf{e}^{(1)}\right] \left[\mathbf{e}^{(1)}\right]'\right) \mathrm{E}\{\mathbf{X}\}\right\}. \tag{3.155}$$

As far as the factor (3.152) is concerned, in order to satisfy (3.148), we shift it by a scalar as follows:

$$F \equiv \left[\mathbf{e}^{(1)}\right]' \mathbf{X} - \left[\mathbf{e}^{(1)}\right]' \mathrm{E}\{\mathbf{X}\}. \tag{3.156}$$

Since this factor is in the form (3.144), we would argue that it represents the PCA factor, when only one factor is required.

It turns out that the above heuristic arguments and conjectures are correct. Furthermore, they can be generalized to any number K of factors. Indeed, the following statements and results hold, see Brillinger (2001).

Consider the $N \times K$ matrix defined as the juxtaposition of the first K eigenvectors:

$$\mathbf{E}_K \equiv \left(\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(K)}\right). \tag{3.157}$$

The solution to the PCA dimension reduction problem (3.147) reads:

$$\{\mathbf{B}_p, \mathbf{A}_p, \mathbf{m}_p\} \equiv \{\mathbf{E}_K, \mathbf{E}_K, (\mathbf{I}_N - \mathbf{E}_K \mathbf{E}_K') \mathrm{E}\{\mathbf{X}\}\}, \tag{3.158}$$

which generalizes (3.155).

The hidden factors that optimally summarize the most information in the invariants by means of affine transformations read:

$$\mathbf{F}_p \equiv \mathbf{E}_K' (\mathbf{X} - \mathrm{E}\{\mathbf{X}\}). \tag{3.159}$$

This expression generalizes (3.156).

From the solution (3.158) we also obtain the expression of the PCA-recovered invariants:

$$\widetilde{\mathbf{X}}_p \equiv \mathrm{E}\{\mathbf{X}\} + \mathbf{E}_K \mathbf{E}_K' (\mathbf{X} - \mathrm{E}\{\mathbf{X}\}). \tag{3.160}$$

This expression generalizes (3.153)-(3.154). As we show in Appendix www.3.5, this expression represents the orthogonal projection of the original invariants onto the hyperplane spanned by the K longest principal axes, i.e. the K-dimensional hyperplane that contains the maximum information about the original invariants, see Figure 3.14.

Furthermore, the perturbations in the PCA dimension reduction model, defined as the residuals $\mathbf{U}_p \equiv \mathbf{X} - \widetilde{\mathbf{X}}_p$, have zero expected value and display zero correlation with the factors:

$$\mathrm{E}\{\mathbf{U}_p\} = \mathbf{0}, \quad \mathrm{Cor}\{\mathbf{F}_p, \mathbf{U}_p\} = \mathbf{0}_{K \times N}, \tag{3.161}$$

see Appendix www.3.5.

Quite obviously, the quality of the approximation provided by the recovered invariants (3.160) depends on the number K of factors. Indeed, we prove in Appendix www.3.5 that the generalized r-square (3.116) can be expressed in terms of the eigenvalues (3.150) of the covariance matrix as follows:

$$R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}_p\right\} = \frac{\sum_{n=1}^{K} \lambda_n}{\sum_{n=1}^{N} \lambda_n}. \tag{3.162}$$

This expression is intuitive. Adding the generic K-th factor to a $(K-1)$-factor PCA analysis corresponds to adding one dimension to the hyperplane

on which the invariants are projected, namely the direction of the K-th largest principal axis of the location-dispersion ellipsoid. On the other hand, the K-th eigenvalue is the variance of the K-th factor:

$$\text{Var}\{F_n\} = \left[\mathbf{e}^{(n)}\right]' \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}' \left[\mathbf{e}^{(n)}\right] = \lambda_n. \tag{3.163}$$

We can thus interpret the K-th eigenvalue as the contribution to the total recovered randomness obtained by adding the K-th dimension of randomness.

In this respect, the numerator in (3.162) is the cumulative contribution to total randomness from the K main dimensions of randomness. Similarly, the denominator is the cumulative contribution to total risk from all the factors, i.e. the denominator represents the total randomness in the invariants.

To summarize, the generalized r-square is the percentage cumulative contribution to total randomness from the K main dimensions of randomness. Notice that the eigenvalues are sorted in decreasing order. Therefore, the marginal contribution of adding one factor decreases with the number of factors.

Idiosyncratic perturbations

Principal component analysis is not the only way to specify the linear hidden-factor model (3.143), which we report here:

$$\mathbf{X} \equiv \mathbf{q} + \mathbf{B}\mathbf{F}(\mathbf{X}) + \mathbf{U}. \tag{3.164}$$

Among other options, one can impose that each of the residual perturbations refer to one and only one invariant, i.e. that the entries of $\mathbf{U}$ be independent of one another.

Imposing this constraint corresponds to factoring the randomness in the market into K contributions common to all the market invariants and N *idiosyncratic perturbations* each of which affects only one invariant.

Nevertheless, the assumption that the perturbations be independent of one another is too strong in general markets. Even the much weaker assumption that the perturbations be uncorrelated is too strong. Indeed this hypothesis, together with the standard assumption (3.118) that factors and perturbations be uncorrelated, is equivalent to the following condition:

$$\text{Cov}\{X_m, X_n\} = \left[\mathbf{B}\,\text{Cov}\{\mathbf{F}(\mathbf{X})\}\,\mathbf{B}'\right]_{mn}, \quad \text{for all } m \neq n. \tag{3.165}$$

This condition can be satisfied in general only in approximation.

Furthermore, the common factors and factor loadings can be identified only modulo an invertible transformation, which we can, but do not have to, assume linear. In other words, if a pair $(\mathbf{F}, \mathbf{B})$ yields a viable model (3.164), so does the pair $\left(\mathbf{AF}, \mathbf{BA}^{-1}\right)$ for any conformable invertible matrix $\mathbf{A}$.

3.4.3 Explicit vs. hidden factors

At this point the legitimate question might arise, whether in order to summarize the randomness in the market is it better to use explicit factors, as discussed in Section 3.4.1, or hidden factors, as discussed in Section 3.4.2.

In general, explicit factor models are easier to interpret, whereas hidden factor models tend to provide a better explanatory power. The first statement is straightforward, therefore we focus on the comparison of the explanatory power of the two methods. Nevertheless, each situation should be evaluated independently.

Consider a generic PCA dimension reduction on the first K factors of an N-dimensional set of invariants $\mathbf{X}$. From (3.160) this process recovers the following invariants:

$$\widetilde{\mathbf{X}}_p \equiv \mathrm{E}\{\mathbf{X}\} + \mathbf{E}_K \mathbf{E}_K' \left(\mathbf{X} - \mathrm{E}\{\mathbf{X}\}\right). \tag{3.166}$$

We recall that the recovered invariants represent the projection of the original invariants onto the K-dimensional hyperplane of maximum randomness spanned by the first K principal axes of the location-dispersion ellipsoid, see Figure 3.15 and compare with Figure 3.14.

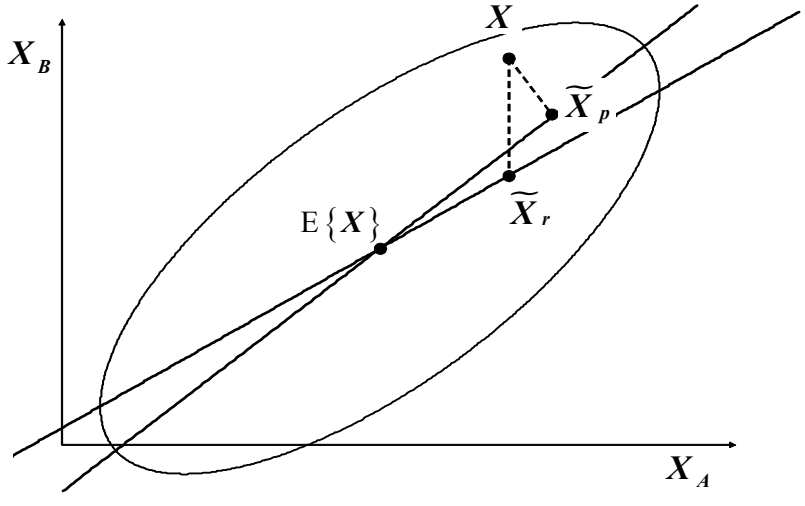

Fig. 3.15. Regression vs. PCA dimension reduction

In order to compare the PCA results with an explicit-factor model we need to restrict our analysis to endogenous explicit-factor models. In other words, first we split the invariants into two subsets:

$$\mathbf{X} \equiv \begin{pmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{pmatrix}, \tag{3.167}$$

where $\mathbf{X}_A$ is a set of K among the N entries of $\mathbf{X}$, and $\mathbf{X}_B$ is the set of the remaining entries. As factors, we consider the variables $\mathbf{X}_A$ and a constant:

$$\mathbf{F} \equiv \begin{pmatrix} 1 \\ \mathbf{X}_A \end{pmatrix}. \tag{3.168}$$

This regression model is completely endogenous, in that the factors are a function of the original invariants. From (3.127) the recovered invariants read:

$$\widetilde{\mathbf{X}}_r \equiv \mathrm{E}\{\mathbf{X}\} + \mathbf{V}_{A,B}\left(\mathbf{X} - \mathrm{E}\{\mathbf{X}\}\right), \tag{3.169}$$

where

$$\mathbf{V}_{A,B} \equiv \left(\begin{array}{c|c} \mathbf{I}_K & \mathbf{0}_{K,N-K} \\ \hline \mathrm{Cov}\{\mathbf{X}_B, \mathbf{X}_A\}\,\mathrm{Cov}\{\mathbf{X}_A\}^{-1} & \mathbf{0}_{N-K,N-K} \end{array}\right). \tag{3.170}$$

Geometrically, the recovered invariants represent the projection of the original invariants along the direction defined by the reference axes of $\mathbf{X}_B$ onto the K-dimensional hyperplane that passes through the expected value and satisfies the following parametric equation:

$$\mathbf{x}_B = \mathrm{E}\{\mathbf{X}_B\} + \mathrm{Cov}\{\mathbf{X}_B, \mathbf{X}_A\}\,\mathrm{Cov}\{\mathbf{X}_A\}^{-1}\left(\mathbf{x}_A - \mathrm{E}\{\mathbf{X}_A\}\right), \tag{3.171}$$

see Figure 3.15. From (3.128) this is the hyperplane that decorrelates the residuals from the factors.

The PCA dimension reduction represents a more symmetrical approach. As such it should yield better results.

Indeed, from (3.147) the PCA approach maximizes the generalized r-square as follows:

$$(\mathbf{G}_p, \mathbf{m}_p) \equiv \underset{(\mathbf{G},\mathbf{m})\in\mathcal{C}_p}{\mathrm{argmax}}\; R^2\{\mathbf{X}, \mathbf{m} + \mathbf{G}\mathbf{X}\}, \tag{3.172}$$

under the only constraint $\mathcal{C}_p$ that the rank of $\mathbf{G}$ be K. This follows because a generic $N \times N$ matrix $\mathbf{G}$ has rank K if and only if it is the product $\mathbf{G} \equiv \mathbf{B}\mathbf{E}'$ of two full-rank $N \times K$ matrices $\mathbf{B}$ and $\mathbf{E}$.

On the other hand, from (3.120) the regression approach maximizes the generalized r-square as follows:

$$(\mathbf{G}_r, \mathbf{m}_r) \equiv \underset{(\mathbf{G},\mathbf{m})\in\mathcal{C}_r}{\mathrm{argmax}}\; R^2\{\mathbf{X}, \mathbf{m} + \mathbf{G}\mathbf{X}\}, \tag{3.173}$$

under a much stronger set of constraints:

$$\mathcal{C}_r : \begin{cases} \mathbf{m}_A \equiv \mathbf{0}_K \\ \mathbf{G}_{AA} \equiv \mathbf{I}_K \\ \mathbf{G}_{AB} \equiv \mathbf{0}_{K,N-K} \\ \mathbf{G}_{BB} \equiv \mathbf{0}_{N-K,N-K}. \end{cases} \tag{3.174}$$

Therefore the PCA approach yields better results:

$$R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}_p\right\} \geq R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}_r\right\}. \tag{3.175}$$

Nevertheless, the regression approach displays other advantages. For instance, the regression dimension reduction is invariant under any rescaling of the factors (3.168), whereas the PCA approach is only invariant under a global rescaling of all the invariants. In practice, one has to be careful to measure the variables in homogenous units when implementing PCA dimension reduction, whereas this is not necessary when implementing regression dimension reduction.

Furthermore, the explanatory power as summarized by the generalized r-square is a statistical identity, whereas the word "explaining" is closely related to the word "understanding": in other words the interpretation of the K invariants $\mathbf{X}_A$ is clear, whereas the interpretation of the K PCA factors might be more obscure.

Finally, if the explanatory variables $\mathbf{X}_A$ are chosen appropriately among the invariants $\mathbf{X}$, regression and principal component analysis yield similar reductions, i.e. (3.175) approaches an equality.

3.4.4 Notable examples

We present here a few notable examples of dimension reduction in the financial markets by means of the techniques discussed in this section: a model for equities, based on one explicit factor and related to the Capital Asset Pricing Model; another model for equity, namely the Fama-French regression, based on three explicit factors; a model for the fixed income market, based on three hidden factors, namely the level-slope-hump PCA decomposition of the yield curve; and a hidden-factor model with idiosyncratic perturbations, related to the Arbitrage Pricing Theory.

Explicit factors and the Capital Asset Pricing Model

Consider a broad stock index like the S&P 500, whose value at the generic time t we denote as M_t. Consider as invariants for a market of N stocks the linear returns (3.10):

$$L_{t,\tau}^{(n)} \equiv \frac{P_t^{(n)}}{P_{t-\tau}^{(n)}} - 1, \quad n = 1, \ldots, N. \tag{3.176}$$

Consider an explicit factors linear model (3.119) based on a constant and one explicit factor, defined as the linear return on the market index:

$$F_{t,\tau}^{M} \equiv \frac{M_t}{M_{t-\tau}} - 1. \tag{3.177}$$

In this case the regression (3.127) recovers the following portion of the stock returns:

$$\widetilde{L}_{t,\tau}^{(n)} \equiv \mathrm{E}\left\{L_{t,\tau}^{(n)}\right\} + \beta_{\tau}^{(n)}\left(F_{t,\tau}^{M} - \mathrm{E}\left\{F_{t,\tau}^{M}\right\}\right), \tag{3.178}$$

where the regression coefficient β is called the *beta* of the stock. From (3.127), the beta is defined as follows:

$$\beta_{\tau}^{(n)} \equiv \frac{\mathrm{Cov}\left\{L_{t,\tau}^{(n)}, F_{t,\tau}^{M}\right\}}{\mathrm{Var}\left\{F_{t,\tau}^{M}\right\}}. \tag{3.179}$$

Notice that the beta depends on the interval τ. Had we used compounded returns as invariants instead, the "square-root rule" (3.76) would have made the beta independent of the interval.

Suppose that the distribution of the linear returns of each stock satisfies the following additional constraint:

$$\mathrm{E}\left\{L_{t,\tau}^{(n)}\right\} = \beta_{\tau}^{(n)}\,\mathrm{E}\left\{F_{t,\tau}^{M}\right\} + \left(1 - \beta_{\tau}^{(n)}\right) R_{t,\tau}^{f}. \tag{3.180}$$

In this expression the *risk-free rate* R is the return on a zero-coupon bond from a period τ before maturity until maturity, which in the notation of Section 3.1.2 reads:

$$R_{t,\tau}^{f} \equiv \left(\frac{1}{Z_{t-\tau}^{(t)}} - 1\right). \tag{3.181}$$

Then the explicit factor model (3.178) becomes the *Capital Asset Pricing Model (CAPM)* of Sharpe (1964), and Lintner (1965), a general equilibrium model for the markets which recovers the following portion of the stock returns:

$$\widetilde{L}_{t,\tau}^{(n)} \equiv R_{t,\tau}^{f} + \beta_{\tau}^{(n)}\left(F_{t,\tau}^{M} - R_{t,\tau}^{f}\right). \tag{3.182}$$

See Ingersoll (1987) for an introduction to the CAPM.

Market-size-type explicit factors

A notable three-factor model for linear returns on stocks is discussed in Fama and French (1993). We consider a set of N stocks, where the generic n-th stock trades at time t at the price $P_t^{(n)}$ and we specify the invariants as the compounded returns (3.11) on these stocks:

$$C_{t,\tau}^{(n)} \equiv \ln\left(\frac{P_t^{(n)}}{P_{t-\tau}^{(n)}}\right). \tag{3.183}$$

The first explicit factor, in addition to a constant, is the compounded return C^M of a broad stock index like the S&P 500. The second factor is the difference SmB ("small minus big") between the compounded returns of a small-cap stock index and the compounded returns of a large-cap stock index; the third factor is the difference HmL ("high minus low") between

the compounded returns of a large book-to-market-value stock index and the compounded returns of a small book-to-market-value stock index. Therefore, this *market-size-type* three-factor linear model reads:

$$\begin{aligned} C_{t,\tau}^{(n)} \equiv & \mathrm{E}\left\{C_{t,\tau}^{(n)}\right\} + \beta^{(n)}\left(C_{t,\tau}^{M} - \mathrm{E}\left\{C_{t,\tau}^{M}\right\}\right) \\ & + \gamma^{(n)}\left(SmB_{t,\tau} - \mathrm{E}\left\{SmB_{t,\tau}\right\}\right) \\ & + \zeta^{(n)}\left(HmL_{t,\tau} - \mathrm{E}\left\{HmL_{t,\tau}\right\}\right) + U_{t,\tau}^{(n)}, \end{aligned} \tag{3.184}$$

where n ranges through all the stocks considered. From (3.127), the regression coefficients (β, γ, ζ) are defined in terms of the cross-covariances among factors and invariants: due to the "square-root" property (3.76), these coefficients do not depend on the estimation interval τ.

Hidden factors and principal component analysis

One of the most widely used applications of hidden factor dimension reduction stems from the principal component analysis of the yield curve. We detail every step of this analysis in our case study, see Section 3.5.2.

Hidden factors and the arbitrage pricing theory model

A notable example of the idiosyncratic approach to hidden factors linear models (3.164) is provided by the *Arbitrage Pricing Theory (APT)* of Ross (1976). Like the CAPM, this is a factor model for the linear returns of the stocks in a broad index such as the S&P 500:

$$\mathbf{L} \equiv \mathrm{E}\{\mathbf{L}\} + \mathbf{B}\mathbf{F}(\mathbf{L}) + \mathbf{U}. \tag{3.185}$$

The APT superimposes a restriction on the distribution of the linear returns, namely:

$$\mathrm{E}\{\mathbf{L}\} = \xi_0 \mathbf{1} + \mathbf{B}\boldsymbol{\xi}, \tag{3.186}$$

where $\mathbf{1}$ is an N-dimensional vector of ones, ξ_0 is a constant and $\boldsymbol{\xi}$ is a K-dimensional vector of risk premia. See Ingersoll (1987) and Connor and Korajczyk (1995) for an introduction to the APT.

3.4.5 A useful routine

In the context of dimension reduction, a challenging problem that often arises is the *selection* of the best K in a pool of N potential candidates to perform a given task. This is a combinatorial problem. The pool of candidates can be indexed by the first N integers:

$$I_N \equiv \{1, \ldots, N\}; \tag{3.187}$$

we have to consider all the possible combinations of K elements from the pool of candidates:

$$I_K \equiv \{n_1, \ldots, n_K\}; \tag{3.188}$$

and we must select the best combination I_K^* among all the above combinations.

For example, consider reducing the dimension by means of an explicit factor model as in Section 3.4.1. There exists a pool of N potential explicit factors:

$$\mathbf{F}_{I_N} \equiv (F_1, \ldots, F_N)', \tag{3.189}$$

but eventually we only consider K among the N potential factors:

$$\mathbf{F}_{I_K^*} \equiv \left(F_{n_1^*}, \ldots, F_{n_K^*}\right)'. \tag{3.190}$$

As another example, consider an allocation problem in a market of N securities, where the final portfolio is constrained to contain a number K of these securities. This dimension-reduction problem is known as *portfolio replication*, namely replicating with as few as K securities a portfolio that should ideally contain N securities.

The best combination I_K^* is defined as the one that maximizes a given objective $\mathcal{O}$:

$$I_K^* = \underset{I_K \subset I_N}{\operatorname{argmax}}\, \mathcal{O}(I_K). \tag{3.191}$$

For instance, in the case of regression dimension reduction the objective is represented by the generalized r-square:

$$\mathcal{O}(I_K) \equiv R^2\left\{\mathbf{X}, \widetilde{\mathbf{X}}(I_K)\right\}, \tag{3.192}$$

see (3.120). In this expression $\widetilde{\mathbf{X}}$ follows from (3.121) and reads:

$$\widetilde{\mathbf{X}}(I_K) \equiv \mathrm{E}\left\{\mathbf{X}\mathbf{F}'_{I_K}\right\}\mathrm{E}\left\{\mathbf{F}_{I_K}\mathbf{F}'_{I_K}\right\}^{-1}\mathbf{F}_{I_K}. \tag{3.193}$$

An alternative specification of the objective is provided for instance by the Akaike criterion, see Parzen, Tanabe, and Kitagawa (1998).

In the case of the PCA approach to dimension reduction the selection problem does not exist, because the PCA factors are naturally sorted in decreasing order of importance, i.e. $I_K^* \equiv (1, \ldots, K)$.

In a portfolio replication problem, the objective is minimizing the tracking error:

$$\mathcal{O}(I_K) \equiv -\,\mathrm{TE}(\boldsymbol{\alpha}(I_K)), \tag{3.194}$$

see (6.179) later in the text.

Combinatorial problems are computationally very challenging. Indeed, the optimization (3.191) implies evaluating the objective $\binom{N}{K}$ times. Furthermore,

the number K is often a decision variable. In other words, the optimal number K is only decided after evaluating the trade-offs of the dimension reduction process, i.e. after computing the following function:

$$K \mapsto \mathcal{O}\left(I_K^*\right), \quad K = 1, \ldots, N. \tag{3.195}$$

For instance, in portfolio replication problems, the ideal number K of securities in the final portfolio is evaluated according to the trade-off between the quality of the replication and the transaction costs.

Computing (3.195) implies evaluating the objective the following number of times:

$$\sum_{K=1}^{N} K \binom{N}{K} = 2^N. \tag{3.196}$$

This number is exorbitant precisely when a dimension reduction is most needed, namely when N is large. Here we propose a routine which evaluates the objective only the following number of times:

$$\sum_{K=1}^{N} K = \frac{N\,(N+1)}{2}. \tag{3.197}$$

The routine proceeds as follows:

Step 0. Set $K \equiv N$, and consider the initial set $I_K \equiv \{1, \ldots, N\}$

Step 1. Consider the K sets obtained from I_K by dropping the generic k-th element:

$$I_K^k \equiv \{n_1, \ldots, n_{k-1}, n_{k+1}, \ldots n_K\}, \quad k = 1, \ldots, K. \tag{3.198}$$

Step 2. Evaluate the above sets:

$$k \mapsto v_K^k \equiv \mathcal{O}\left(I_K^k\right), \quad k = 1, \ldots, K. \tag{3.199}$$

Step 3. Determine the worst element in I_K:

$$k^* \equiv \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \left\{v_K^k\right\}. \tag{3.200}$$

Step 4. Drop the worst element in I_K:

$$I_{K-1} \equiv I_K^{k^*}. \tag{3.201}$$

Step 5. If $K = 2$ stop. Otherwise set $K \equiv K - 1$ and go to Step 1.

Although this routine yields suboptimal results, in practice it proved very close to optimal in a variety of applications. In other words, the function

$$K \mapsto v_K^{k^*}, \quad K = 1, \ldots, N. \tag{3.202}$$

is in general a very good approximation of (3.195).

3.5 Case study: modeling the swap market

In this section we discuss how to model the swap market. Swaps are very liquid securities and many new contracts are traded every day. A *v-swap* $(E-t)$-*forward* is a contract whose value at the generic time t reads:

$$P_t^{(E,v)} \equiv s\rho \sum_{k=1}^{v/\rho} Z_t^{(E_k)} + Z_t^{(E+v)} - Z_t^{(E)}. \tag{3.203}$$

In this formula s is the agreed upon *fixed rate* expressed in annualized terms: at inception t_0 this rate is typically set in such a way that the value of the contract zero, i.e. it is set as the $(E-t_0)$-into-v forward par swap rate defined in (3.57); ρ is a fixed time-interval of the order of a few months; the generic term $E_k \equiv E + k\rho$ is one *fixed leg payment date*; $Z_t^{(E)}$ is the price of a zero coupon bond with maturity E. The pricing formula (3.203) originates from the structure of the contract, according to which agreed upon fixed payments are swapped against floating payments that depend on the current levels of interest rates, see Rebonato (1998) and Brigo and Mercurio (2001). Nevertheless, we can take (3.203) as the definition of a security.

In this case study the investment decision is taken at $T \equiv$ January 1st 2000 and we plan to invest in an "eight year swap two years forward", i.e. a swap that starts $(E-T) \equiv$ two years from the investment date on $E \equiv$ January 1st 2002 and ends $v \equiv$ eight years later on $E+v \equiv$ January 1st 2010. The fixed payments occur every $\rho \equiv$ three months. Therefore, this contract is determined by the price of thirty-three zero-coupon bonds.

We assume that the investment horizon is $\tau \equiv$ two months. Our aim is to determine the distribution of $P_{T+\tau}^{(E,v)}$. To do this, we dispose of the daily database of all the zero-coupon bond prices for the past ten years.

3.5.1 The market invariants

Everyday, many new forward swap contracts are issued with new starting and ending dates. Therefore, the swap market is completely priced by the set of all the zero coupon bond prices for virtually all the maturities on a daily basis up to around thirty years in the future:

$$Z_t^{(E)} \text{ such that } E = t + 1d, t + 2d, \ldots, t + 30y. \tag{3.204}$$

The first step to model a market is to determine its invariants. We have seen in Section 3.1.2 that the natural invariants for the fixed income market are the changes in yield to maturity:

$$X_{t,\tilde{\tau}}^{(v)} \equiv Y_t^{(v)} - Y_{t-\tilde{\tau}}^{(v)}. \tag{3.205}$$

In this expression $\tilde{\tau}$ is the estimation interval and v denotes a specific time to maturity in the *yield curve*, which is the plot of the yield to maturity as a function of the respective time to maturity: