4

Estimating the distribution of the market invariants

In this chapter we discuss how to estimate the distribution of the market invariants from empirical observations.

In Section 4.1 we define the concept of estimator, which is simply a function of current information that yields a number, the estimate. Such a general definition includes estimators that perform poorly, i.e. functions that yield an estimate which has little in common with the real distribution of the market invariants. Therefore we discuss optimality criteria to evaluate an estimator.

After defining estimators and how to evaluate them, we need to actually construct estimators for the market invariants. Nevertheless, constructing estimators by maximizing the above optimality criteria is not possible. First of all, the search of the best estimator among all possible functions of current information is not feasible. Secondly the optimality criteria rarely yield a univocal answer. In other words, an estimator might perform better than another one in given circumstances, and worse in different circumstances. Therefore, we construct estimators from general intuitive principles, making sure later that their performance is acceptable, and possibly improving them with marginal corrections. In this spirit, we proceed as follows.

In Section 4.2 we introduce nonparametric estimators. These estimators are based on the law of large numbers. Therefore, they perform well when the number of empirical observations in the time series of the market invariants is large, see Figure 4.1. When this is the case, nonparametric estimators are very flexible, in that they yield sensible estimates no matter the underlying true distribution of the market invariants. In particular we discuss the sample quantile, the sample mean, the sample covariance and the ordinary least square estimate of the regression factor loadings in an explicit factor model, stressing the geometrical properties of these estimators. We conclude with an overview of kernel estimators.

When the number of observations is not very large, nonparametric estimators are no longer suitable. Therefore we take a parametric approach, by assuming that the true distribution of the market invariants belongs to a restricted class of potential distributions. In Section 4.3 we discuss maximum

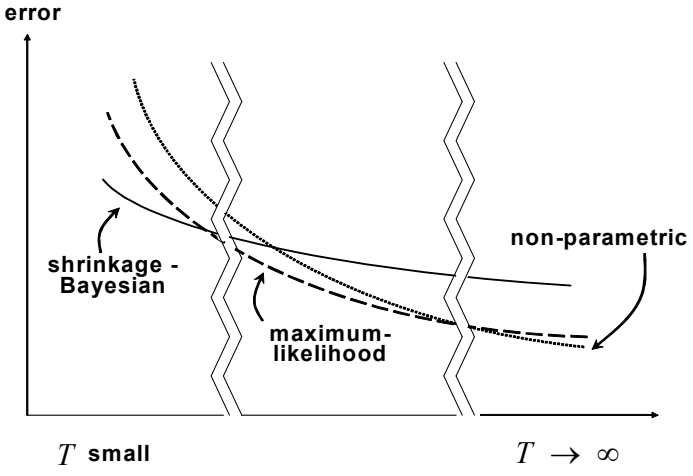

Fig. 4.1. Performance of different types of estimators

likelihood estimators, which are built in such a way that the past observations of the market invariants become the most likely outcomes of the estimated parametric distribution. We compute the maximum likelihood estimators of location, dispersion, and factor loadings under the assumption that the market invariants are elliptically distributed: this shows the intrinsic outlier-rejection mechanism of maximum likelihood estimators. Then we study thoroughly the normal case: as it turns out, the main driver of the performance of the maximum likelihood estimators is the overall level of correlation among the market invariants, as summarized by the condition number.

In some applications the number of observations is so scanty, and the result of the estimate so unreliable, that it is advisable to average the final estimates with fixed, yet potentially wrong, values: this way we obtain shrinkage estimators, see Figure 4.1. In Section 4.4 we discuss the shrinkage estimators for the location parameters, the dispersion parameters and the factor loadings of a linear model.

In Section 4.5 we discuss robust estimation. Indeed, the parametric approach dramatically restricts the set of potential distributions for the market invariants. Robust estimation provides a set of techniques to evaluate and possibly fix the consequences of not having included the true, unknown distribution among the set of potential distributions. We discuss classical measures of robustness, such as the sensitivity curve and the jackknife, and develop the more general concept of influence function. Then we evaluate the robustness of the estimators previously introduced in this chapter and show how to build robust M-estimators. In particular, we discuss M-estimators of the

location parameters, the dispersion parameters and the factor loadings of a linear model.

In Section 4.6 we conclude with a series of practical tips to improve the estimation of the distribution of the market invariants in specific situations. Among other issues, we discuss outliers detection, which we tackle by means of high breakdown estimators such as the minimum volume ellipsoid and the minimum covariance determinant; missing data, which we tackle by means of the EM algorithm; and weighted estimation techniques such as the exponential smoothing, which accounts for the higher reliability of more recent data with respect to data farther back in the past.

4.1 Estimators

Before introducing the concept of estimator, we review our working assumptions, which we set forth in Section 3.1.

The randomness in the market is driven by the market invariants. The invariants are random variables that refer to a specific estimation-horizon $\widetilde{\tau}$ and are *independent and identically distributed (i.i.d.)* across time. The generic invariant $\mathbf{X}_{t,\widetilde{\tau}}$ becomes known at the respective time t, which is part of the set of equally spaced estimation dates:

$$t \in \mathcal{D}_{\widetilde{t},\widetilde{\tau}} \equiv \left\{\widetilde{t}, \widetilde{t}+\widetilde{\tau}, \widetilde{t}+2\widetilde{\tau}, \ldots\right\}. \tag{4.1}$$

For example, we have seen in Section 3.1.1 that the invariants in the equity market are the compounded returns. In other words, for a stock that at the generic time t trades at the price P_t, the following set of random variables are independent and identically distributed across time, as t varies in (4.1):

$$X_{t,\widetilde{\tau}} \equiv \ln\left(\frac{P_t}{P_{t-\widetilde{\tau}}}\right), \quad t \in \mathcal{D}_{\widetilde{t},\widetilde{\tau}}. \tag{4.2}$$

Furthermore, these variables become known at time t.

Notice that once the time origin $\widetilde{t}$ and the time interval $\widetilde{\tau}$ have been fixed, we can measure time in units of $\widetilde{\tau}$ and set the origin in $\widetilde{t}-\widetilde{\tau}$. This way, without loss of generality, we can always reduce the estimation dates to the set of positive integers:

$$t \in \mathcal{D}_{\widetilde{t},\widetilde{\tau}} \equiv \{1, 2, 3, \ldots\}. \tag{4.3}$$

We will use this more convenient notation throughout this chapter.

In this notation the market invariants of our example, namely the compounded returns (4.2), read:

$$X_t \equiv \ln\left(\frac{P_t}{P_{t-1}}\right), \quad t = 1, 2, \ldots. \tag{4.4}$$

Since the invariants are independent and identically distributed across time, from (2.44) we obtain that their across-time joint distribution, as represented by their probability density function, factors as follows:

$$f_{\mathbf{X}_1,\mathbf{X}_2,\ldots}(\mathbf{x}_1, \mathbf{x}_2, \ldots) = f_{\mathbf{X}}(\mathbf{x}_1) f_{\mathbf{X}}(\mathbf{x}_2) \cdots, \tag{4.5}$$

where we stress that the single-period joint distribution of the invariants $f_{\mathbf{X}}$ does not depend on the time index. Therefore all the information about the invariants is contained in one single-period multivariate distribution.

In (4.5) we chose to represent the distribution of the invariants in terms of their probability density function $f_{\mathbf{X}}$. Equivalently, we might find it more convenient to represent the distribution of the invariants in terms of either the cumulative distribution function $F_{\mathbf{X}}$ or the characteristic function $\phi_{\mathbf{X}}$, see Figure 2.2. The factorization (4.5) holds for any representation, as we see from (2.46) and (2.48).

4.1.1 Definition

Our aim is to infer the single-period distribution of the invariants. More precisely, we aim at inferring the "truth", as represented by a generic number S of features of the distribution of the market invariants. These features can be expressed as an S-dimensional vector of functionals of the probability density function (or of the cumulative distribution function, or of the characteristic function):

$$\mathbf{G}[f_{\mathbf{X}}] \equiv \text{"unknown truth"}. \tag{4.6}$$

For example, if we are interested in the expected value of the compounded return on a stock (4.4), the "unknown truth" is the following one-dimensional functional:

$$G[f_X] \equiv \int_{-\infty}^{+\infty} x f_X(x)\, dx, \tag{4.7}$$

where f_X is the unknown probability density function of any compounded return X_t, and does not depend on the specific time t.

The current time is T. We base inference on the information i_T about the invariants available at the time T when the investment decision is made. This information is represented by the time series of all the past realizations of the invariants:

$$i_T \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}, \tag{4.8}$$

where the lower-case notation stresses the fact that the once random variables $\mathbf{X}_t$ have become observable numbers. An *estimator* is a vector-valued

function that associates a vector in $\mathbb{R}^S$, i.e. a set of S numbers, with available information:

$$\boxed{\textbf{estimator: information } i_T \mapsto \textbf{number } \widehat{\mathbf{G}}} \tag{4.9}$$

For example the following is an estimator:

$$\widehat{G}[i_T] \equiv \frac{1}{T}\sum_{t=1}^{T} x_t. \tag{4.10}$$

Notice that the definition of estimator (4.9) is not related to the goal of estimation (4.6). Again, an estimator is simply a function of currently available information.

For example, the following is a function of information and thus it is an estimator:

$$\widehat{G}[i_T] \equiv x_1 x_T. \tag{4.11}$$

Similarly, for strange that it might sound, the following is also an estimator:

$$\widehat{G}[i_T] \equiv 3. \tag{4.12}$$

4.1.2 Evaluation

Although the definition of estimator is very general, an estimator serves its purpose only if its value is close to the true, unknown value (4.6) that we are interested in:

$$\widehat{\mathbf{G}}[i_T] \approx \mathbf{G}[f_{\mathbf{X}}]. \tag{4.13}$$

To make this statement precise, we need a criterion to evaluate estimators. In order to evaluate an estimator, the main requirement is its *replicability*: an estimator is good not only if the result of the estimation is close to the true unknown value, but also if this does not happen by chance.

For example, the estimator (4.12) could yield by chance the true, unknown parameter if this happens to be equal to 3, much like the hands of a broken watch happen to display the correct time twice a day.

To tackle replicability, notice that the available information (4.8), namely the time series of the market invariants, is the realization of a set of random variables:

$$I_T \equiv \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}. \tag{4.14}$$

In a different scenario, the realization of this set of variables would have assumed a different value i'_T and therefore the outcome of the estimate would

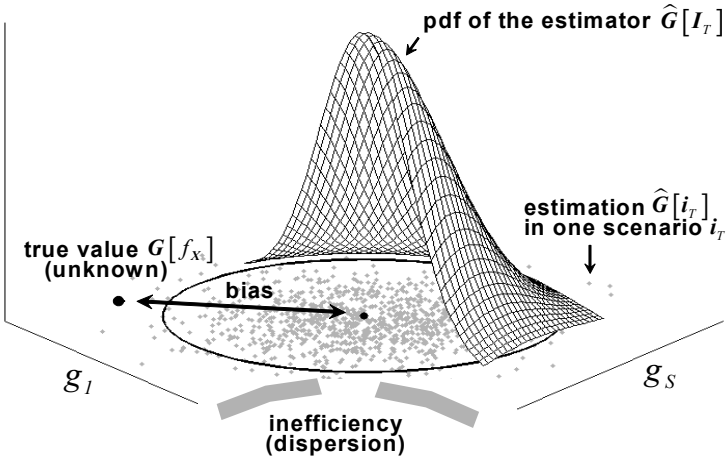

Fig. 4.2. Estimation: replicability, bias and inefficiency

have been a different number $\widehat{\mathbf{G}}[i'_T]$, see Figure 4.2. Therefore the estimator (4.9), as a function of the random variable I_T instead of the specific occurrence i_T, becomes a (multivariate) *random variable*:

$$\widehat{\mathbf{G}}[i_T] \mapsto \widehat{\mathbf{G}}[I_T]. \tag{4.15}$$

The distribution of the information (4.14) is fully determined by the true, unknown distribution $f_{\mathbf{X}}$ of the market invariants through (4.5). Therefore, the distribution of the estimator (4.15) is also determined by the true, unknown distribution $f_{\mathbf{X}}$ of the market invariants, see Figure 4.2.

For example, if the invariants (4.4) are normally distributed with the following unknown parameters:

$$X_t \sim \mathrm{N}\left(\mu, \sigma^2\right), \tag{4.16}$$

then the estimator (4.10) is normally distributed with the following parameters:

$$\widehat{G}[I_T] \equiv \frac{1}{T}\sum_{t=1}^{T} X_t \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{T}\right), \tag{4.17}$$

where μ and σ^2 are unknown.

The distribution associated with an estimator is at least as important as the specific outcome $\widehat{\mathbf{G}}[i_T]$ of the estimation process: an estimator is suitable, i.e. (4.13) holds, if the distribution of the multivariate random variable $\widehat{\mathbf{G}}[I_T]$

is highly concentrated around the true unknown value $\mathbf{G}[f_{\mathbf{X}}]$. For instance, this is not the case in Figure 4.2.

Suppose we use the estimator (4.10) to estimate (4.7), i.e. the expected value of the invariants. From (4.16) this reads:

$$G[f_X] = \mu. \tag{4.18}$$

Therefore the distribution of the estimator (4.17) is centered around the true unknown value μ and the concentration of this distribution is of the order of $\sigma/\sqrt{T}$.

Nevertheless, evaluating a multivariate distribution can be complex. To summarize the goodness of an estimator into a univariate distribution we introduce the *loss*:

$$\text{Loss}\left(\widehat{\mathbf{G}}, \mathbf{G}\right) \equiv \left\|\widehat{\mathbf{G}}[I_T] - \mathbf{G}[f_{\mathbf{X}}]\right\|^2, \tag{4.19}$$

where $\|\cdot\|$ denotes a norm, see $(A.7)$. For reasons to become clear in a moment, it is common to induce the norm from a quadratic form, i.e. a symmetric and positive $S \times S$ matrix $\mathbf{Q}$ such that the following relation holds true:

$$\|\mathbf{v}\|^2 \equiv \mathbf{v}'\mathbf{Q}\mathbf{v}. \tag{4.20}$$

Since the loss is the square of a norm, from $(A.7)$ the loss is zero only for those outcomes where the estimator $\widehat{\mathbf{G}}$ yields an estimate that is equal to the true value to be estimated, and is strictly positive otherwise. Therefore, the estimator is good if the distribution of the loss is tightly squeezed above the value of zero.

In our example, from (4.17) and (4.18) we obtain:

$$\widehat{G}[I_T] - G[f_X] \sim \text{N}\left(0, \frac{\sigma^2}{T}\right). \tag{4.21}$$

We can summarize the goodness of this estimator with the quadratic loss induced by $Q \equiv 1$ in (4.20). Then from (1.106) we obtain the distribution of the loss, which is the following central gamma with one degree of freedom:

$$\text{Loss}\left(\widehat{G}, G\right) \equiv \left(\widehat{G} - G\right)^2 \sim \text{Ga}\left(1, \frac{\sigma^2}{T}\right). \tag{4.22}$$

In the presence of a large number of observations, or when the underlying market is not too volatile, this loss is a random variable tightly squeezed above the value of zero.

Even evaluating the shape of a univariate distribution can be complex, see Chapter 5. To further summarize the analysis of the goodness of an estimator

we consider the expected value of the loss: the higher the expected value, the worse the performance of the estimator. Since the loss is a square distance, we consider the square root of the expectation of the loss. The *error*[1] is the average distance between the outcome of the estimation process and the true value to be estimated over all the possible scenarios:

$$\mathrm{Err}\left(\widehat{\mathbf{G}}, \mathbf{G}\right) \equiv \sqrt{\mathrm{E}\left\{\left\|\widehat{\mathbf{G}}\left(I_T\right) - \mathbf{G}\left[f_{\mathbf{X}}\right]\right\|^2\right\}}. \tag{4.23}$$

In our example, from (4.22) and (1.113) the error reads:

$$\mathrm{Err}\left(\widehat{G}, G\right) = \frac{\sigma}{\sqrt{T}}. \tag{4.24}$$

As expected, the larger the number of observations in the time series and the lower the volatility of the market, the lower the estimation error.

The definition (4.19)-(4.20) of the loss in terms of a square norm and the definition (4.23) of the error as the square root of its expected value are not the only viable choices. Nevertheless, the above definitions are particularly intuitive because they allow to decompose the error into bias and inefficiency.

The *bias* measures the distance between the "center" of the distribution of the estimator and the true unknown parameter to estimate:

$$\mathrm{Bias}^2\left[\widehat{\mathbf{G}}, \mathbf{G}\right] \equiv \left\|\mathrm{E}\left\{\widehat{\mathbf{G}}\left[I_T\right]\right\} - \mathbf{G}\left[f_{\mathbf{X}}\right]\right\|^2, \tag{4.25}$$

see Figure 4.2.

The *inefficiency* is a measure of the dispersion of the estimator, and as such it does not depend on the true unknown value:

$$\mathrm{Inef}^2\left[\widehat{\mathbf{G}}\right] \equiv \mathrm{E}\left\{\left\|\widehat{\mathbf{G}}\left[I_T\right] - \mathrm{E}\left\{\widehat{\mathbf{G}}\left[I_T\right]\right\}\right\|^2\right\}, \tag{4.26}$$

see Figure 4.2.

It is easy to check that in terms of bias and inefficiency the error (4.23) factors as follows:

$$\mathrm{Err}^2\left[\widehat{\mathbf{G}}, \mathbf{G}\right] = \mathrm{Bias}^2\left[\widehat{\mathbf{G}}, \mathbf{G}\right] + \mathrm{Inef}^2\left[\widehat{\mathbf{G}}\right]. \tag{4.27}$$

In these terms, the statement that the replicability distribution of a good estimator is highly peaked around the true value can be rephrased as follows: a good estimator is very efficient and displays little bias.

[1] The error is called *risk* in the statistical literature. We prefer to reserve this term for financial risk

In our example, from (4.17) and (4.18) we obtain the bias:

$$\text{Bias}\left[\widehat{G}, G\right] = \left|\text{E}\left\{\widehat{G}\left[I_T\right]\right\} - G\left[f_X\right]\right| = |\mu - \mu| = 0. \tag{4.28}$$

In other words, the estimator is centered around the true, unknown value μ.

From (4.17) we obtain the inefficiency:

$$\text{Inef}\left[\widehat{G}\right] = \text{Sd}\left\{\widehat{G}\left[I_T\right]\right\} = \frac{\sigma}{\sqrt{T}}. \tag{4.29}$$

In other words, the estimator has a dispersion of the order of $\sigma/\sqrt{T}$.

Comparing with (4.24) we see that the factorization (4.27) holds.

Notice that the definitions of loss and error are scale dependent: for example if the true value $\mathbf{G}$ has the dimension of money and we measure it in US dollars, the error is about one hundred times smaller than if we measure it in Japanese yen. To make the evaluation scale independent we can normalize the loss and the error by the length of the true value, if this length is not zero. Therefore at times we consider the percentage loss, which is a random variable:

$$\text{PLoss}\left(\widehat{\mathbf{G}}, \mathbf{G}\right) \equiv \frac{\left\|\widehat{\mathbf{G}}\left(I_T\right) - \mathbf{G}\left[f_{\mathbf{X}}\right]\right\|^2}{\left\|\mathbf{G}\left[f_{\mathbf{X}}\right]\right\|^2}; \tag{4.30}$$

and the percentage error, which is a scale-independent number:

$$\text{PErr}\left(\widehat{\mathbf{G}}, \mathbf{G}\right) \equiv \frac{\sqrt{\text{E}\left\{\left\|\widehat{\mathbf{G}}\left(I_T\right) - \mathbf{G}\left[f_{\mathbf{X}}\right]\right\|^2\right\}}}{\left\|\mathbf{G}\left[f_{\mathbf{X}}\right]\right\|}. \tag{4.31}$$

An estimator is suitable if its percentage error is much smaller than one.

At this point we face a major problem: the distribution of the loss, and thus the error of an estimator, depends on the underlying true distribution of the market invariants $f_{\mathbf{X}}$. If this distribution were known, we would not need an estimator in the first place.

In our example, from (4.24) the error of the estimator (4.10) depends on the standard deviation σ of the unknown distribution of the invariants (4.16): this estimator is good if the invariants are not too volatile.

Similarly, the estimator (4.12) gives rise to a deterministic loss, which is equal to the error and reads:

$$\text{Err}\left(\widehat{G}, G\right) = \left|\mu - 3\right|. \tag{4.32}$$

This estimator is suitable if the expected value of the invariants happens to lie in the neighborhood of the value $\mu \equiv 3$.

Nevertheless, neither μ nor σ are known parameters.

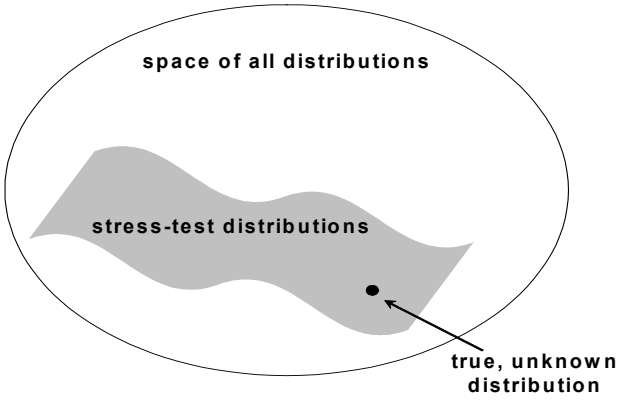

Fig. 4.3. Evaluation of estimators: choice of stress-test distributions

Therefore in order to evaluate an estimator we have to proceed as follows.

First we consider, among all the possible distributions of the market invariants, a subset of stress test distributions that is large enough to contain the true, unknown distribution, see Figure 4.3.

Then we make sure that the estimator is suitable, i.e. its distribution is peaked around the true unknown value to be estimated for all the distributions in the stress test set, see Figure 4.4.

In general an estimator performs well with some stress test distributions and performs poorly with other stress test distributions, see Figure 4.4. Consequently, in choosing the set of stress test distributions we face the following dichotomy: on the one hand, the stress test set should be as broad as possible, in such a way to encompass the true, unknown distribution; on the other hand, the stress test set should be as narrow as possible, in such a way that estimators can be built which display small errors for all the stress test distributions.

4.2 Nonparametric estimators

Assume that the number of observations T in the time series i_T is very large. The nonparametric approach is based on the following intuitive result, well known to practitioners: under fairly general conditions, sample averages computed over the whole time series approximate the expectation computed with the true distribution, and the approximation improves with the number of observations in the time series.

This result is known as the *law of large numbers (LLN)*, which we represent as follows:

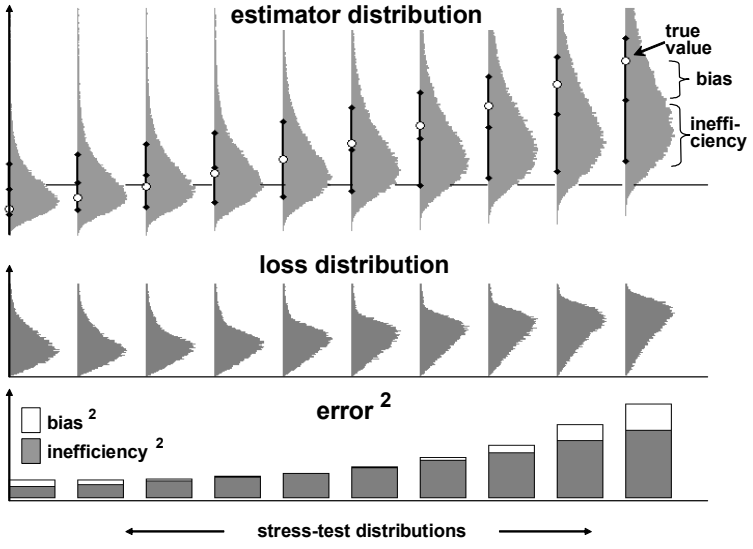

Fig. 4.4. Evaluation of estimators: loss and error

$$\frac{1}{T}\sum_{t=1}^{T}\{\text{past}\} \underset{T\to\infty}{\approx} \mathrm{E}\{\text{future}\}. \tag{4.33}$$

The Law of Large Numbers implies the *Glivenko-Cantelli theorem*. This theorem states that the empirical distribution (2.239) of a set of independent and identically distributed variables, as represented for example by its cumulative distribution function, tends[2] to the true distribution as the number of observations goes to infinity, see Figure 4.5:

$$\lim_{T\to\infty} F_{i_T}(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x}). \tag{4.34}$$

Expression (4.34) suggests how to define the estimator of a generic functional $\mathbf{G}[f_{\mathbf{X}}]$ of the true, yet unknown, distribution of the market invariants. Indeed, we only need to replace in the functional $\mathbf{G}[f_{\mathbf{X}}]$ the true, unknown probability density function $f_{\mathbf{X}}$ with the empirical probability density function (2.240), which we report here:

$$f_{i_T}(\mathbf{x}) \equiv \frac{1}{T}\sum_{t=1}^{T}\delta^{(\mathbf{x}_t)}(\mathbf{x}), \tag{4.35}$$

where δ is the Dirac delta ($B.17$). In other words we define the estimator of $\mathbf{G}[f_{\mathbf{X}}]$ as follows:

[2] One should specify the topology for the limits in the law of large numbers and in the Glivenko-Cantelli theorem, see e.g. Shirayaev (1989) for details. Here we choose a heurisitc approach.

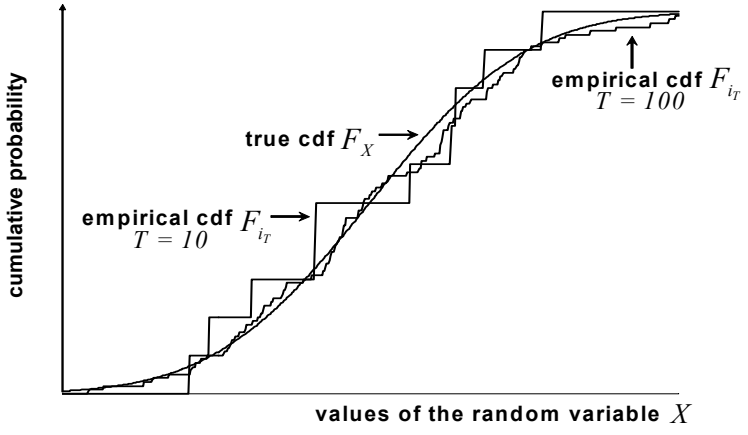

Fig. 4.5. Glivenko-Cantelli theorem

$$\widehat{\mathbf{G}}[i_T] \equiv \mathbf{G}[f_{i_T}]. \tag{4.36}$$

To test the goodness of this estimator we should compute its replicability, i.e. the distribution of $\widehat{G}[I_T]$ as in (4.15), for all possible distributions. This is an impossible task. Nevertheless, under fairly general conditions, when the number of observations T is very large the *central limit theorem (CLT)* states that the estimator is approximately normally distributed:

$$\widehat{\mathbf{G}}[I_T] \sim \mathrm{N}\left(\mathbf{G}[f_{\mathbf{X}}], \frac{\mathbf{A}}{T}\right), \tag{4.37}$$

where $\mathbf{A}$ is a suitable symmetric and positive matrix [3]. The above approximation becomes exact only in the limit of an infinite number of observations T in the time series: although this limit is never attained in practice, for a large enough number of observations the nonparametric approach yields benchmark estimators that can subsequently be refined.

We now use the nonparametric approach to estimate the features of the distribution of the market invariants that are most interesting in view of financial applications.

[3] The matrix $\mathbf{A}$ is defined in terms of the influence function (4.185) as follows:

$$A_{jk} \equiv \int_{\mathbb{R}^N} \mathrm{IF}(\mathbf{x}, f_{\mathbf{X}}, G_j)\, \mathrm{IF}(\mathbf{x}, f_{\mathbf{X}}, G_k)\, f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x},$$

see Huber (1981).

4.2.1 Location, dispersion and hidden factors

If the invariants X_t are univariate random variables, we can use as location parameter the generic quantile q_p, which is defined implicitly in terms of the probability density function f_X of the invariant as follows:

$$\int_{-\infty}^{q_p[f_X]} f_X(x)\,dx \equiv p, \tag{4.38}$$

see (1.18). By applying (4.36) to the definition of quantile, we obtain the respective estimator $\widehat{q}_p$. This is the *sample quantile* (1.124):

$$\widehat{q}_p[i_T] \equiv x_{[pT]:T}, \tag{4.39}$$

where $[\cdot]$ denotes the integer part. In particular, for $p \equiv 1/2$ this expression becomes the *sample median*.

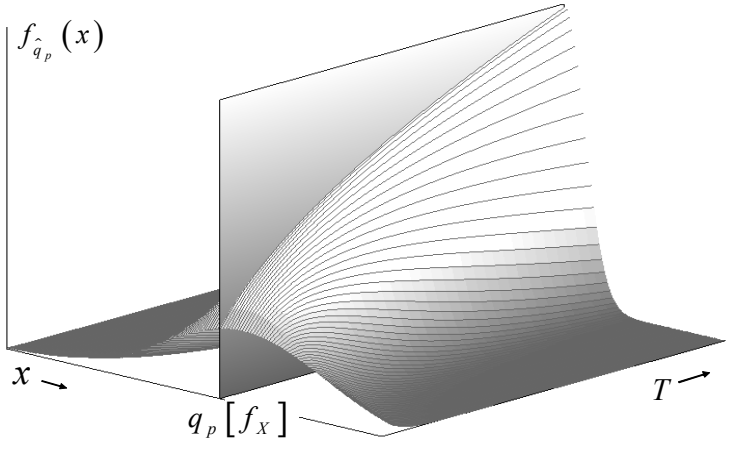

Fig. 4.6. Sample quantile: evaluation

To evaluate this estimator, we consider it as a random variable as in (4.15). From (2.248) the probability density function of the estimator $\widehat{q}_p$ reads:

$$f_{\widehat{q}_p}(x) = \frac{T!\,[F_X(x)]^{[pT]-1}\,[1-F_X(x)]^{T-[pT]}\,f_X(x)}{([pT]-1)!\,(T-[pT])!}. \tag{4.40}$$

From (2.253) this density is concentrated around the quantile q_p and from (2.252) the quality of the estimator improves as the sample size T increases, see Figure 4.6.

Similarly, to estimate the dispersion of the univariate invariants X_t we can use the *sample interquantile range*, derived by applying (4.36) to (1.37).

In the multivariate case, we can rely on the expected value of the invariant $\mathbf{X}$ as parameter of location. We derive the nonparametric estimator of the expected value by applying (4.36) to the definition (2.54) of expected value. This is the expected value of the empirical distribution (2.244), i.e. the *sample mean*:

$$\widehat{\mathrm{E}}\left[i_T\right] \equiv \int_{\mathbb{R}^N} \mathbf{x} f_{i_T}\left(\mathbf{x}\right) d\mathbf{x} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t. \tag{4.41}$$

Similarly, as a multivariate parameter of dispersion we choose the covariance matrix. By applying (4.36) to the definition (2.67) of covariance we derive the respective nonparametric estimator:

$$\widehat{\mathrm{Cov}}\left[i_T\right] = \frac{1}{T}\sum_{t=1}^{T} \left(\mathbf{x}_t - \widehat{\mathrm{E}}\left[i_T\right]\right)\left(\mathbf{x}_t - \widehat{\mathrm{E}}\left[i_T\right]\right)'. \tag{4.42}$$

This is the covariance matrix of the empirical distribution (2.245), i.e. the *sample covariance*.

From (4.42) we derive an expression for the estimator of the principal component decomposition of the covariance matrix. Indeed, it suffices to compute the PCA decomposition of the sample covariance:

$$\widehat{\mathrm{Cov}} \equiv \widehat{\mathbf{E}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{E}}'. \tag{4.43}$$

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

$$\widehat{\mathbf{\Lambda}} \equiv \mathrm{diag}\left(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_N\right); \tag{4.44}$$

and $\widehat{\mathbf{E}}$ is the orthogonal matrix of the respective sample eigenvectors. The matrix $\widehat{\mathbf{E}}$ is the estimator of the PCA factor loadings, and the entries of $\widehat{\mathbf{\Lambda}}$ are the estimators of the variances of the PCA factors.

We do not evaluate here the performance of the estimators (4.41), (4.42), (4.43) and (4.44) on a set of stress test distributions, because the same estimators reappear in a different context in Section 4.3.

The sample mean and the sample covariance display an interesting geometrical interpretation. To introduce this property, consider a generic N-dimensional vector $\boldsymbol{\mu}$ and a generic $N \times N$ scatter matrix $\boldsymbol{\Sigma}$, i.e. a symmetric and positive matrix. Consider the ellipsoid ($A.73$) defined by these two parameters, see Figure 4.7:

$$\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \equiv \left\{\mathbf{x} \in \mathbb{R}^N \text{ such that } \left(\mathbf{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}-\boldsymbol{\mu}\right) \leq 1\right\}. \tag{4.45}$$

Consider now the set of the Mahalanobis distances from $\boldsymbol{\mu}$ through the metric $\boldsymbol{\Sigma}$ of each observation $\mathbf{x}_t$ in the time series of the invariants:

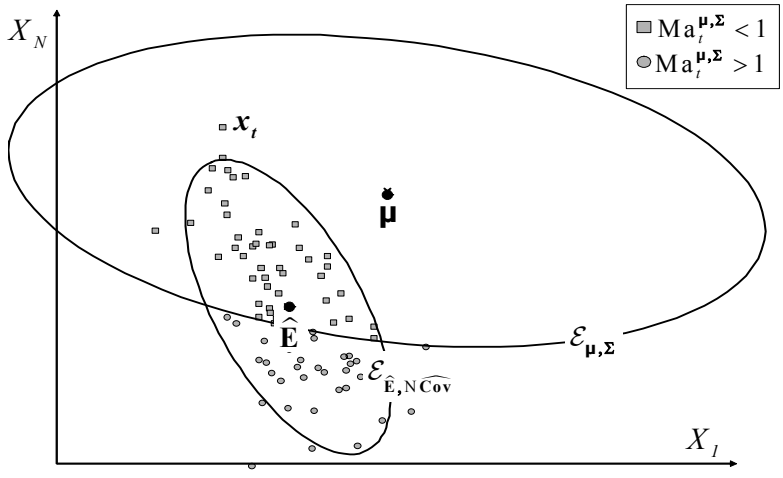

Fig. 4.7. Sample mean and sample covariance: geometric properties

$$\mathrm{Ma}_t^{\boldsymbol{\mu},\boldsymbol{\Sigma}} \equiv \mathrm{Ma}\left(\mathbf{x}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \sqrt{\left(\mathbf{x}_t - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_t - \boldsymbol{\mu}\right)}. \tag{4.46}$$

The Mahalanobis distance is the "radius" of the ellipsoid concentric to $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ that crosses the observation $\mathbf{x}_t$. In particular, if $\mathrm{Ma}_t^{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ is one, then the observation $\mathbf{x}_t$ lies on the ellipsoid (4.45). Consider the average of the square distances:

$$\overline{r^2}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \frac{1}{T} \sum_{t=1}^{T} \left(\mathrm{Ma}_t^{\boldsymbol{\mu},\boldsymbol{\Sigma}}\right)^2. \tag{4.47}$$

If this number is close to one, the ellipsoid passes through the cloud of observations.

The sample mean and sample covariance represent the choices of location and scatter parameter respectively that give rise to the smallest ellipsoid among all those that pass through the cloud of observations, see Figure 4.7. More formally, we prove in Appendix www.4.1 the following result:

$$\left(\widehat{\mathrm{E}}, N\widehat{\mathrm{Cov}}\right) = \underset{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathcal{C}}{\mathrm{argmin}} \left[\mathrm{Vol}\left\{\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\right\}\right], \tag{4.48}$$

where the set of constraints $\mathcal{C}$ imposes that $\boldsymbol{\Sigma}$ be symmetric and positive and that the average Mahalanobis distance be one:

$$\overline{r^2}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv 1. \tag{4.49}$$

In other words, the set of constraints $\mathcal{C}$ imposes that the respective ellipsoid (4.45) passes trough the cloud of observations, see Figure 4.7.

The result (4.48) is intuitive: the ellipsoid generated by the sample mean and covariance is the one that best fits the observations, since all the observations are packed in its neighborhood.

Nevertheless, in some circumstances the ellipsoid $\mathcal{E}_{\widehat{\mathrm{E}},N\widehat{\mathrm{Cov}}}$ "tries too hard" to embrace all the observations: if an observation is an outlier, the sample mean and the sample covariance tend to perform rather poorly in an effort to account for this single observation. We discuss this phenomenon further in Section 4.5.

4.2.2 Explicit factors

Consider the explicit factor affine model (3.119), which we report here:

$$\mathbf{X} = \mathbf{B}\mathbf{F} + \mathbf{U}. \tag{4.50}$$

Since we observe both the N-dimensional market invariants $\mathbf{X}$ and the K-dimensional explicit factors $\mathbf{F}$, the available information (4.8) consists of the time series of both the invariants and the factors:

$$i_T \equiv \{\mathbf{x}_1, \mathbf{f}_1, \ldots, \mathbf{x}_T, \mathbf{f}_T\}. \tag{4.51}$$

By applying (4.36) to the definition of the regression factor loadings (3.121) we obtain the nonparametric estimator of the regression factor loadings of the explicit factor affine model:

$$\widehat{\mathbf{B}}[i_T] \equiv \left(\sum_t \mathbf{x}_t \mathbf{f}_t'\right)\left(\sum_t \mathbf{f}_t \mathbf{f}_t'\right)^{-1}. \tag{4.52}$$

This matrix represents the *ordinary least square (OLS)* estimator of the regression factor loadings.

The name is due to a geometric property of the OLS coefficients, which we sketch in Figure 4.8. Indeed, as we show in Appendix www.4.1, the OLS estimator $\widehat{\mathbf{B}}$ provides the best fit to the observations, in the sense that it minimizes the sum of the square distances between the original observations $\mathbf{x}_t$ and the recovered values $\widehat{\mathbf{B}}\mathbf{f}_t$:

$$\widehat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}} \sum_t \|\mathbf{x}_t - \mathbf{B}\mathbf{f}_t\|^2, \tag{4.53}$$

where $\|\cdot\|$ is the standard norm (A.6).

By applying (4.36) to the covariance of the residuals (3.129) we obtain the respective nonparametric estimator:

$$\widehat{\mathrm{Cov}}[i_T] \equiv \frac{1}{T}\sum_t \left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)\left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)'. \tag{4.54}$$

This is the *ordinary least square (OLS)* estimator of the covariance of the residuals.

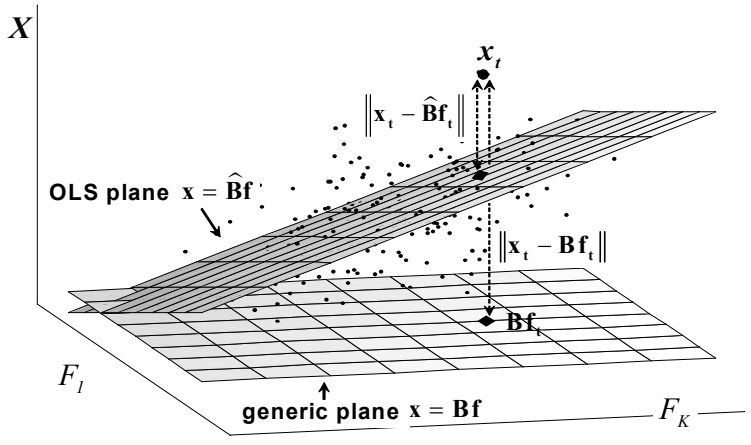

Fig. 4.8. OLS estimates of factor loadings: geometric properties

4.2.3 Kernel estimators

Here we briefly put into perspective a nonparametric approach to estimation that is becoming very popular in financial applications, see Campbell, Lo, and MacKinlay (1997).

The nonparametric estimators defined by the recipe (4.36) are very sensitive to the input data and thus are not robust, in a sense to be discussed precisely in Section 4.5. Intuitively, this happens because the empirical probability density function (4.35), is a sum of Dirac deltas, which are not regular, smooth functions.

One way to solve this problem consists in replacing the empirical distribution with a regularized, smoother distribution by means of the convolution, see $(B.54)$. In other words, we replace the empirical probability density function as follows:

$$f_{i_T} \mapsto f_{i_T;\epsilon} \equiv \frac{1}{T}\sum_{t=1}^{T} \frac{1}{(2\pi)^{\frac{N}{2}}\,\epsilon^N} e^{-\frac{1}{2\epsilon^2}(\mathbf{x}-\mathbf{x}_t)'(\mathbf{x}-\mathbf{x}_t)}. \tag{4.55}$$

The outcome of this operation is a smoother empirical probability density function such as the one sketched in Figure 2.18. In this context, the Gaussian exponential, or any other smoothing function, takes the name of *kernel*, and the width ϵ of the regularizing function takes on the name of *bandwidth*.

Once the probability density function has been smoothened, we can define new nonparametric estimators that replace (4.36) as follows:

$$\widehat{\mathbf{G}}[i_T] \equiv \mathbf{G}[f_{i_T;\epsilon}]. \tag{4.56}$$

The bandwidth of the kernel must be chosen according to the following trade-off. A narrow bandwidth gives rise to non-robust estimators: indeed, a null bandwidth gives rise to the benchmark estimators (4.36) stemming from the non-regularized empirical distribution. On the other hand, a wide bandwidth blends the data too much and gives rise to loss of information.

4.3 Maximum likelihood estimators

In this section we abandon the nonparametric approach. In the parametric approach the stress test set of potential distributions, which include the true, unknown distribution of the market invariants, is dramatically restricted. Only a few models of distributions are considered: once the model is chosen, it is subsequently fitted to the empirical data.

We represent a parametric the family of potential distributions, the stress test distributions, in terms of their probability density function $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is an S-dimensional parameter that fully determines the distribution and that ranges in a given set $\boldsymbol{\Theta}$, see Figure 4.9 and compare with Figure 4.3.

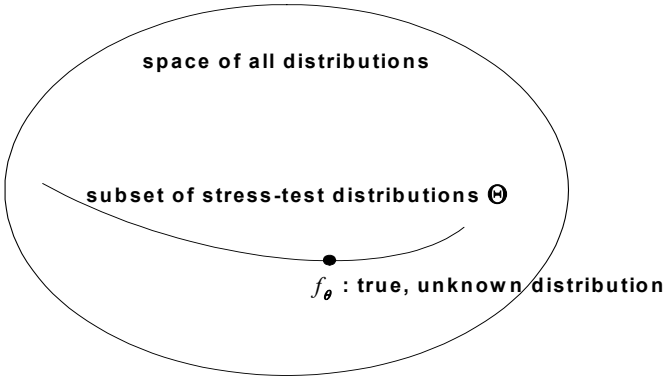

Fig. 4.9. Parametric approach to estimation

For example, from empirical observations it might seem reasonable to model a given market invariant by means of the lognormal distribution as follows:

$$X_t \sim \mathrm{LogN}(\theta, 1). \tag{4.57}$$

In this case the distribution's parameters are one-dimensional; the distribution's probability density function reads:

$$f_\theta(x) \equiv \frac{1}{\sqrt{2\pi}x} e^{-\frac{1}{2}(\ln x - \theta)^2}; \tag{4.58}$$

and the parameter space Θ is the real line $\mathbb{R}$.

Since the distribution of the invariants is completely determined by the parameters $\boldsymbol{\theta}$, estimating the distribution corresponds to determining these parameters. In other words, the estimation process (4.9) consists of determining some function of the available information $\widehat{\boldsymbol{\theta}}[i_T]$ that is close to the true, unknown parameters.

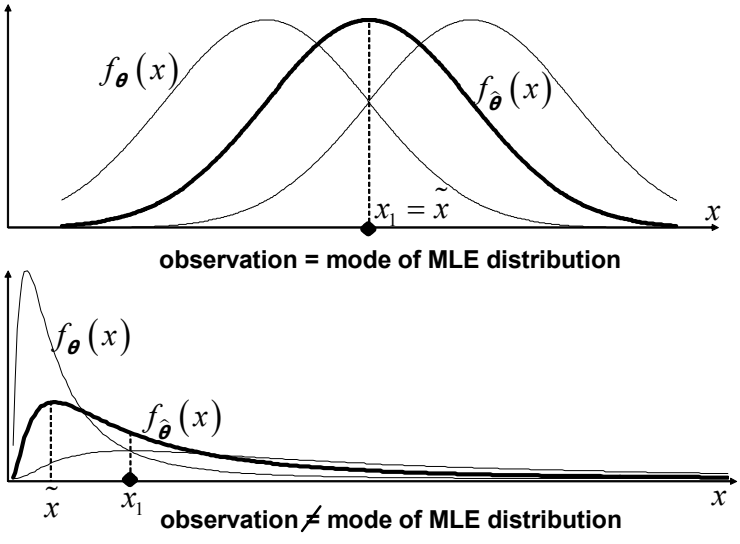

Fig. 4.10. Maximum likelihood estimator as mode

The *maximum likelihood principle* provides a method to determine an estimator which is related to the intuitive concept of mode. We recall that the mode $\widetilde{\mathbf{x}}$ of a distribution $f_{\mathbf{X}}$ is the value that corresponds to the peak of the distribution, i.e. the largest value of the probability density function:

$$\widetilde{\mathbf{x}} \equiv \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmax}} f_{\mathbf{X}}(\mathbf{x}). \tag{4.59}$$

Suppose that only one observation $\mathbf{x}_1$ is available. Most likely, this observation lies in a region where the probability density function is comparatively large, i.e. near the the mode. Therefore, once we assume that the distribution that generated that observation belongs to a specific parametric family $f_{\mathbf{X}} \equiv f_{\boldsymbol{\theta}}$, the most intuitive value for the parameter $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ that makes the pdf in that point the largest, see the top plot in Figure 4.10.

In other words, according to the maximum likelihood principle we define the estimator $\widehat{\boldsymbol{\theta}}$ as follows:

$$\widehat{\boldsymbol{\theta}} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} f_{\boldsymbol{\theta}}(\mathbf{x}_1). \tag{4.60}$$

Notice that, although the maximum likelihood estimator draws on the concept of mode, the observation $\mathbf{x}_1$ does not necessarily turn out to be the mode of the distribution $f_{\widehat{\boldsymbol{\theta}}}$:

$$\mathbf{x}_1 \neq \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmax}} f_{\widehat{\boldsymbol{\theta}}}(\mathbf{x}), \tag{4.61}$$

see the bottom plot in Figure 4.10.

For example, from (4.58) we solve:

$$\widehat{\theta} \equiv \underset{\theta\in\mathbb{R}}{\operatorname{argmax}} \left\{ \frac{1}{\sqrt{2\pi}x_1} e^{-\frac{1}{2}(\ln x_1 - \theta)^2} \right\}. \tag{4.62}$$

From the first-order condition with respect to θ we obtain the value:

$$\widehat{\theta} = \ln x_1. \tag{4.63}$$

On the other hand, from the first-order condition with respect to x we obtain that the mode $\widehat{x}$ satisfies:

$$\ln \widehat{x} = \widehat{\theta} - 1. \tag{4.64}$$

Therefore, in the case of the lognormal distribution, (4.61) takes place, i.e. the mode of the distribution estimated with the maximum likelihood principle is not the observation, see the bottom plot in Figure 4.10.

In the general case of a time series of several observations, from (4.5) we obtain the joint probability density function of the time series, which is the product of the single-period probability density functions:

$$f_{\boldsymbol{\theta}}(i_T) \equiv f_{\boldsymbol{\theta}}(\mathbf{x}_1) \cdots f_{\boldsymbol{\theta}}(\mathbf{x}_T). \tag{4.65}$$

Expression (4.65) is also called the *likelihood function* of the time series. Now we can apply the maximum likelihood principle (4.60) to the whole time series. Therefore the *maximum likelihood estimator (MLE)* of the parameters $\boldsymbol{\theta}$ is defined as follows:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}[i_T] &\equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} f_{\boldsymbol{\theta}}(i_T) \\ &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{t=1}^{T} \ln f_{\boldsymbol{\theta}}(\mathbf{x}_t). \end{aligned} \tag{4.66}$$

For example, in the case of lognormal invariants, from (4.58) we solve:

$$\widehat{\theta} \equiv \underset{\theta\in\Theta}{\operatorname{argmax}} \left\{ -\sum_{t=1}^{T} \frac{1}{2} \left(\ln x_t - \theta\right)^2 \right\}. \tag{4.67}$$

The first-order condition reads:

$$0 = \frac{1}{T} \sum_{t=1}^{T} \left(\ln x_t - \widehat{\theta}\right), \tag{4.68}$$

which implies the following expression for the maximum likelihood estimate of the parameter:

$$\widehat{\theta} = \frac{1}{T} \sum_{t=1}^{T} \ln x_t. \tag{4.69}$$

The maximum likelihood estimator displays a few appealing properties.

For instance, the *invariance property*, which states that the MLE of a function of the parameters is that function applied to the MLE of the parameters:

$$\widehat{g(\boldsymbol{\theta})} = g\left(\widehat{\boldsymbol{\theta}}\right). \tag{4.70}$$

This property follows from the definition (4.66).

Furthermore, similarly to the nonparametric approach (4.37), the maximum likelihood principle provides good estimators in the limit case of a very large number of observations T in the time series i_T, as sketched in Figure 4.1. Indeed, the following relation holds in approximation, and the approximation becomes exact as T tends to infinity:

$$\widehat{\boldsymbol{\theta}}\left[i_T\right] \sim \mathrm{N}\left(\boldsymbol{\theta}, \frac{\boldsymbol{\Gamma}}{T}\right). \tag{4.71}$$

In this expression $\boldsymbol{\Gamma}$ is a symmetric and positive matrix called the *Fisher information matrix*:

$$\boldsymbol{\Gamma} \equiv \mathrm{Cov}\left\{ \frac{\partial \ln\left(f_{\boldsymbol{\theta}}\left(\mathbf{X}\right)\right)}{\partial \boldsymbol{\theta}} \right\}, \tag{4.72}$$

see e.g. Haerdle and Simar (2003).

The *Cramer-Rao lower bound* theorem states that the inefficiency of the maximum likelihood estimator, as represented by (4.72), is the smallest possible achievable with an unbiased estimator, and from (4.71) we see that the MLE becomes unbiased in the limit of many observations.

Nevertheless, we introduced the parametric approach to estimation in order to build estimators that perform well in the realistic case of a finite number of observations of market invariants. Therefore below we evaluate the maximum likelihood estimators of parametric models that are apt to describe the market invariants.

4.3.1 Location, dispersion and hidden factors

In Chapter 3 we saw that the market invariants are quite symmetrical. Therefore, in this section we construct and evaluate maximum-likelihood estimators under the assumption that the N-dimensional invariants $\mathbf{X}$ are elliptically distributed:

$$\mathbf{X} \sim \mathrm{El}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g\right), \tag{4.73}$$

where $\boldsymbol{\mu}$ is the N-dimensional location parameter, $\boldsymbol{\Sigma}$ is the $N \times N$ dispersion matrix and g is the probability density generator, see (2.268). In other words, the probability density function of the invariants invariants $\mathbf{X}$ is of the form:

$$f_{\boldsymbol{\theta}}\left(\mathbf{x}\right) \equiv \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} g\left(\mathrm{Ma}^2\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right), \tag{4.74}$$

where $\mathrm{Ma}\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the *Mahalanobis distance* of the point $\mathbf{x}$ from the point $\boldsymbol{\mu}$ through the metric $\boldsymbol{\Sigma}$:

$$\mathrm{Ma}\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \sqrt{\left(\mathbf{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}-\boldsymbol{\mu}\right)}, \tag{4.75}$$

see (2.61).

Under the assumption (4.73) the parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ completely determine the distribution of the market invariants[4]. These parameters span the set:

$$\boldsymbol{\Theta} \equiv \left\{\boldsymbol{\mu} \in \mathbb{R}^N, \boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}, \boldsymbol{\Sigma} \succeq \mathbf{0}\right\}, \tag{4.76}$$

where $\succeq \mathbf{0}$ denotes symmetric and positive.

In this context, estimating the distribution of the market invariants means estimating from currently available information i_T the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In Appendix www.4.2 we prove that the MLE estimators $\widehat{\boldsymbol{\mu}}\left[i_T\right]$ and $\widehat{\boldsymbol{\Sigma}}\left[i_T\right]$ are the solutions to the following joint set of implicit equations:

$$\widehat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \frac{w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)}{\sum_{s=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_s, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)} \mathbf{x}_t \tag{4.77}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)' w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right), \tag{4.78}$$

where the function w is defined as follows in terms of the probability density generator:

$$w\left(z\right) \equiv -2\frac{g'\left(z\right)}{g\left(z\right)}. \tag{4.79}$$

Notice that defining the following weights:

[4] We assume known the specific density generator, otherwise we would obtain a *semiparametric* model.

$$w_t \equiv w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right), \tag{4.80}$$

we can interpret the maximum likelihood estimators of location and dispersion (4.77) and (4.78) as weighted sums:

$$\widehat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \frac{w_t}{\sum_{s=1}^{T} w_s} \mathbf{x}_t \tag{4.81}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_{t=1}^{T} w_t \left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)'. \tag{4.82}$$

Each observation is weighted according to its Mahalanobis distance from the ML estimator of location through the metric defined by the ML estimator of dispersion.

For example, assume that the market invariants are Cauchy distributed, see (2.208). In this case the density generator reads:

$$g^{\mathrm{Ca}}(z) = \frac{\Gamma\left(\frac{1+N}{2}\right)}{\Gamma\left(\frac{1}{2}\right)(\pi)^{\frac{N}{2}}} (1+z)^{-\frac{1+N}{2}}, \tag{4.83}$$

where Γ is the gamma function (B.80). Therefore the weights (4.80) become:

$$w_t = \frac{N+1}{1+\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)}. \tag{4.84}$$

This is a decreasing function of the Mahalanobis distance: the maximum likelihood estimators of location and dispersion of a set of Cauchy-distributed invariants tend to neglect outliers.

This result is intuitive: we recall that the Cauchy distribution is fat-tailed, see Figure 1.9. Therefore extreme observations, i.e. observations with large Mahalanobis distance, are quite frequent. These extreme observations might distort the estimation, which is why the maximum likelihood estimator tends to taper their influence in the estimation process.

After solving (4.77)-(4.79) for $\widehat{\boldsymbol{\Sigma}}$ we can derive the expression for the maximum likelihood estimator of the principal component factor model. Indeed, it suffices to compute the PCA decomposition of the estimator:

$$\widehat{\boldsymbol{\Sigma}}[i_T] \equiv \widehat{\mathbf{E}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{E}}', \tag{4.85}$$

where $\widehat{\boldsymbol{\Lambda}}$ is the diagonal matrix of the eigenvalues in decreasing order and $\widehat{\mathbf{E}}$ is the orthogonal matrix of the respective eigenvectors. Then $\widehat{\mathbf{E}}$ becomes the MLE estimator of the hidden factor loadings and $\widehat{\boldsymbol{\Lambda}}$ becomes the estimator of the dispersion of the hidden factors.

To evaluate the performance of the maximum likelihood estimators of location and dispersion $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ we should determine the distribution of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ when in (4.77)-(4.79) the market invariants are considered as random variables as in (4.15).

Unfortunately, in the generic elliptical case it is not possible to convert the implicit equations (4.77)-(4.79) into explicit functional expressions of current information. Therefore we must solve for the estimators numerically and resort to simulations to evaluate their performance, unless the invariants are normally distributed. We discuss the normal case in detail in Section 4.3.3.

4.3.2 Explicit factors

Consider the explicit factor affine model (3.119), which we report here:

$$\mathbf{X} = \mathbf{B}\mathbf{F} + \mathbf{U}. \tag{4.86}$$

Since we observe both the N-dimensional invariants $\mathbf{X}$ and the K-dimensional factors $\mathbf{F}$, the available information (4.8) is the time series of both the invariants and the factors:

$$i_T \equiv \{\mathbf{x}_1, \mathbf{f}_1, \ldots, \mathbf{x}_T, \mathbf{f}_T\}. \tag{4.87}$$

To implement the maximum likelihood approach we could model the $(N+K)$-dimensional joint distribution of invariants and factors by means of some parametric distribution $f_{\boldsymbol{\theta}}$ and then maximize the likelihood over the parameters $\boldsymbol{\theta}$ and the factor loadings $\mathbf{B}$.

Nevertheless, most explicit factor models serve the purpose of stress testing the behavior of the invariants under assumptions on the future realization of the factors. For example, practitioners ask themselves such questions as what happens to a given stock if the market goes up, say, 2%. Therefore, it is more convenient to model the N-dimensional distribution of the perturbations $f_{\boldsymbol{\theta}|\mathbf{f}} \equiv f_{\mathbf{U}|\mathbf{f}}$ *conditional* on knowledge of the factors and model the *conditional* distribution of the invariants accordingly:

$$\mathbf{X}|\mathbf{f} = \mathbf{B}\mathbf{f} + \mathbf{U}|\mathbf{f}. \tag{4.88}$$

Under the above assumptions the *conditional* distribution $f_{\mathbf{X}|\mathbf{f}}$ of the invariants becomes a parametric function $f_{\boldsymbol{\theta},\mathbf{B}}$ of the parameters $\boldsymbol{\theta}$ of the perturbations and the factor loadings $\mathbf{B}$. Therefore we can apply the maximum likelihood principle (4.66) to the conditional distribution of the invariants, determining the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of the distribution of the perturbations and the maximum likelihood estimator of the factor loadings $\widehat{\mathbf{B}}$:

$$\left(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{B}}\right) \equiv \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\mathbf{B}} f_{\boldsymbol{\theta},\mathbf{B}}(i_T). \tag{4.89}$$

In Chapter 3 we saw that the market invariants are quite symmetrical. Therefore, we construct the maximum-likelihood estimators under the assumption that the conditional distribution of the perturbations be an N-dimensional elliptical random variable:

$$\mathbf{U}_t|\mathbf{f}_t \sim \mathrm{El}\left(\mathbf{0}, \boldsymbol{\Sigma}, g\right). \tag{4.90}$$

In other words we assume that the perturbations are centered in zero; that $\boldsymbol{\Sigma}$ is their $N \times N$ dispersion matrix and that g is their probability density generator.

From (2.270) the invariants are elliptically distributed with the same generator:

$$\mathbf{X}_t|\mathbf{f}_t \sim \mathrm{El}\left(\mathbf{B}\mathbf{f}_t, \boldsymbol{\Sigma}, g\right). \tag{4.91}$$

In this context the parameters to be estimated are $\mathbf{B}$ and $\boldsymbol{\Sigma}$.

In Appendix www.4.2 we show that the MLE estimators of these parameters solve the following set of joint implicit equations:

$$\widehat{\mathbf{B}} = \left[\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\mathbf{x}_t\mathbf{f}_t'\right] \left[\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\mathbf{f}_t\mathbf{f}_t'\right]^{-1} \tag{4.92}$$

and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)\left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)', \tag{4.93}$$

where the function w is defined in terms of the probability density generator:

$$w\left(z\right) \equiv -2\frac{g'\left(z\right)}{g\left(z\right)}. \tag{4.94}$$

In the generic elliptical case, the implicit equations (4.92)-(4.94) must be solved numerically and the evaluation of the estimators must be performed by means of simulations. On the other hand, in the specific normal case the above implicit equations can be solved analytically. We discuss the normal explicit factor model at the end of Section 4.3.3.

4.3.3 The normal case

In the special case where the market invariants are normally distributed the analysis of the maximum likelihood estimators of location, dispersion and explicit factors can be performed analytically. This analysis provides insight into the more general case.

Location, dispersion and hidden factors

Assume that the market invariants are normally distributed:

$$\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.95}$$

In the normal case the location parameter $\boldsymbol{\mu}$ is the expected value of the distribution and the dispersion parameter $\boldsymbol{\Sigma}$ is its covariance matrix.

The normal distribution is a special case of elliptical distribution, which corresponds to the following choice of the density generator:

$$g^{\mathrm{N}}(z) \equiv \frac{e^{-\frac{z}{2}}}{(2\pi)^{\frac{N}{2}}}, \tag{4.96}$$

see (2.264). It is immediate to check that in the normal case the weights (4.79) are constant:

$$w(z) \equiv 1. \tag{4.97}$$

To interpret this result, we compare it with the respective result for the Cauchy distribution. The normal distribution is very thin-tailed and therefore extreme observations are rare. If an observation is far from the location parameter, the reason must be due to a large dispersion matrix: therefore, unlike (4.84), the maximum likelihood estimator gives full weight to that observation, in such a way to effectively modify the estimation and lead to a larger estimate of the dispersion matrix.

From (4.77) we obtain the explicit expression of the estimator of location in terms of current information:

$$\widehat{\boldsymbol{\mu}}[i_T] = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t. \tag{4.98}$$

Similarly, from (4.78) we obtain the explicit expression of the estimator of dispersion in terms of current information:

$$\widehat{\boldsymbol{\Sigma}}[i_T] = \frac{1}{T}\sum_{t=1}^{T} (\mathbf{x}_t - \widehat{\boldsymbol{\mu}})(\mathbf{x}_t - \widehat{\boldsymbol{\mu}})'. \tag{4.99}$$

These estimators are the sample mean (4.41) and the sample covariance (4.42) respectively. It is reassuring that two completely different methods yield the same estimators for both location and dispersion. This supports our statement that the sample mean and the sample covariance are the benchmark estimators of location and dispersion respectively.

To evaluate the goodness of the sample mean and of the sample covariance under the normal hypothesis we proceed as in (4.15), computing the joint distribution of the following random variables:

$$\widehat{\boldsymbol{\mu}}[I_T] \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_t \tag{4.100}$$

$$\widehat{\boldsymbol{\Sigma}}[I_T] \equiv \frac{1}{T}\sum_{t=1}^{T} (\mathbf{X}_t - \widehat{\boldsymbol{\mu}})(\mathbf{X}_t - \widehat{\boldsymbol{\mu}})'. \tag{4.101}$$

In Appendix www.4.3 we prove the following results. The sample mean is normally distributed:

$$\widehat{\boldsymbol{\mu}}\,[I_T] \sim \mathrm{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{T}\right). \tag{4.102}$$

The distribution of the sample covariance is related to the Wishart-distribution (2.223) by the following expression:

$$T\widehat{\boldsymbol{\Sigma}}\,[I_T] \sim \mathrm{W}\,(T-1, \boldsymbol{\Sigma}). \tag{4.103}$$

Furthermore, (4.102) and (4.103) are independent of each other.

- Component-wise evaluation

From the above expressions we can evaluate component-wise the error (4.23) of the sample estimators, using the standard quadratic form $Q \equiv 1$ in (4.20) and decomposing the error into bias and inefficiency as in (4.27).

For the sample mean, from (4.102) we obtain:

$$\mathrm{Bias}\,(\widehat{\mu}_i, \mu_i) = 0 \tag{4.104}$$

$$\mathrm{Inef}\,(\widehat{\mu}_i) = \sqrt{\frac{\Sigma_{ii}}{T}}. \tag{4.105}$$

This shows that the sample mean is unbiased and that its inefficiency shrinks to zero as the number of observations grows to infinity.

As for the estimator of the sample covariance, from (4.103) and (2.227)-(2.228) we obtain:

$$\mathrm{Bias}\left(\widehat{\Sigma}_{mn}, \Sigma_{mn}\right) = \frac{1}{T}\,|\Sigma_{mn}| \tag{4.106}$$

$$\mathrm{Inef}\left(\widehat{\Sigma}_{mn}\right) = \sqrt{\frac{T-1}{T^2}}\sqrt{\Sigma_{mm}\Sigma_{nn} + \Sigma_{mn}^2}. \tag{4.107}$$

As expected, bias and inefficiency shrink to zero as the number of observations grows to infinity.

Formulas (4.104)-(4.107) provide the measure of performance for each of the entries of the estimators separately. It is nonetheless interesting to obtain a global measure of performance. Since the sample mean $\widehat{\boldsymbol{\mu}}$ and the sample covariance $\widehat{\boldsymbol{\Sigma}}$ are independent, we evaluate them separately.

- Evaluation of sample mean

To evaluate the sample mean (4.100), we consider the loss (4.19) induced by the quadratic form $\mathbf{Q} \equiv \mathbf{I}_N$. In other words, the loss is the following random variable:

$$\mathrm{Loss}\,(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \equiv [\widehat{\boldsymbol{\mu}}\,[I_T] - \boldsymbol{\mu}]'\,[\widehat{\boldsymbol{\mu}}\,[I_T] - \boldsymbol{\mu}]. \tag{4.108}$$

We then summarize the information contained in the loss by means of the error (4.23). We prove in Appendix www.4.3 that the error reads:

$$\mathrm{Err}^2\left(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}\right) = \frac{1}{T}\mathrm{tr}\left(\boldsymbol{\Sigma}\right). \tag{4.109}$$

The whole error is due to inefficiency, as the sample estimator is unbiased:

$$\mathrm{Inef}^2\left(\widehat{\boldsymbol{\mu}}\right) = \frac{1}{T}\mathrm{tr}\left(\boldsymbol{\Sigma}\right) \tag{4.110}$$

$$\mathrm{Bias}^2\left(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}\right) = 0. \tag{4.111}$$

As expected, the error decreases as the number of observations grows to infinity. Furthermore, it is an increasing function of the average variance: intuitively, more volatile invariants give rise to larger estimation errors.

To gain further insight into the estimation error of the sample mean, we consider the PCA decomposition ($A.70$) of the scatter matrix:

$$\boldsymbol{\Sigma} \equiv \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'. \tag{4.112}$$

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

$$\boldsymbol{\Lambda} \equiv \mathrm{diag}\left(\lambda_1, \ldots, \lambda_N\right); \tag{4.113}$$

and $\mathbf{E}$ is the juxtaposition of the respective orthogonal eigenvectors. From the PCA decomposition the following identity follows:

$$\mathrm{tr}\left[\boldsymbol{\Sigma}\right] = \mathrm{tr}\left[\boldsymbol{\Lambda}\right]. \tag{4.114}$$

Therefore, the estimation error of sample mean (4.109), along with its factorization in terms of bias and inefficiency, is completely determined by the eigenvalues of $\boldsymbol{\Sigma}$. To interpret this result geometrically, consider the ellipsoid $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ determined by the market parameters as described in ($A.73$), which is also the location-dispersion ellipsoid of the invariants (2.75). Since the eigenvalues represent the (square of) the length of the principal axes of the ellipsoid, the estimation error of the sample mean is completely determined by the shape of the location-dispersion ellipsoid of the invariants, and not by its location or orientation.

In particular, a key parameter is the *condition number* or the *condition ratio* defined as the ratio between the smallest and the largest eigenvalue:

$$\mathrm{CN}\left\{\mathbf{X}\right\} \equiv \frac{\lambda_N}{\lambda_1}. \tag{4.115}$$

The condition number ranges in the interval $[0, 1]$. When the condition number is close to one the invariants $\mathbf{X}$ are *well-conditioned* and the location-dispersion ellipsoid that represents the invariants resembles a sphere. When the condition number is close to zero the invariants $\mathbf{X}$ are *ill-conditioned*: the ellipsoid is elongated, shaped like a cigar, since the actual dimension of risk is less than the number of invariants. This is the case in highly correlated markets, such as the swap market, see Figure 3.20.

To capture the effect of the shape of the location-dispersion ellipsoid on the estimation error, we keep the location $\boldsymbol{\mu}$ constant and we let the scatter matrix $\boldsymbol{\Sigma}$ vary as follows:

$$\boldsymbol{\Sigma} \equiv \begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \theta \\ \theta & \cdots & \theta & 1 \end{pmatrix}, \quad \theta \in (0,1). \tag{4.116}$$

The parameter θ represents the overall level of correlation among the invariants: as the correlation varies between zero and one, the condition number varies between one and zero.

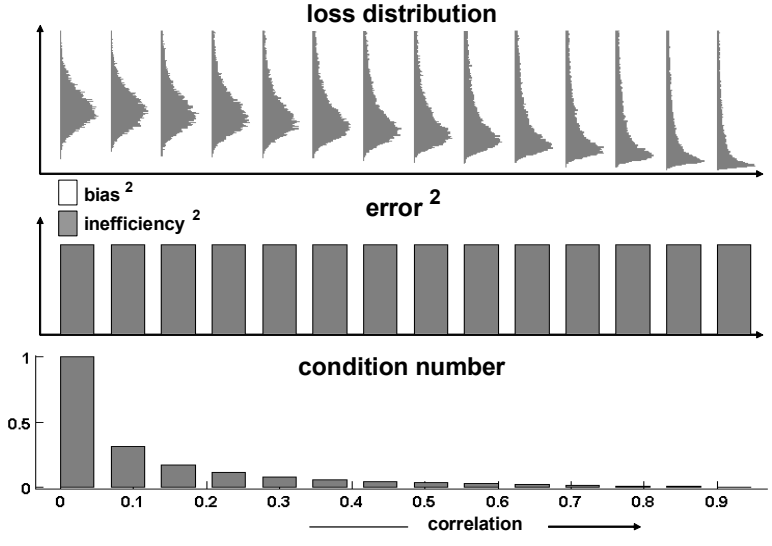

Fig. 4.11. Sample mean: evaluation

In Figure 4.11 we display the distribution of the loss (4.108) and the respective error (4.109) as the market parameters vary according to (4.116). Notice how the distribution of the loss varies, although the inefficiency and thus the error remain constant.

- Evaluation of sample covariance

To evaluate the sample covariance (4.101) we introduce the *Frobenius quadratic form* for a generic symmetric matrix $\mathbf{S}$:

$$\|\mathbf{S}\|^2 \equiv \mathrm{tr}\left[\mathbf{S}^2\right]. \tag{4.117}$$

This corresponds to the choice $\mathbf{Q} \equiv \mathbf{I}_{N^2}$ in (4.20) acting on vec $(\mathbf{S})$, the stacked columns of $\mathbf{S}$. Accordingly, the loss (4.19) becomes the following random variable:

$$\text{Loss}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) \equiv \text{tr}\left[\left(\widehat{\boldsymbol{\Sigma}}\left[I_T\right] - \boldsymbol{\Sigma}\right)^2\right]. \tag{4.118}$$

In Appendix www.4.3 we show that the estimation error (4.23) relative to this loss reads:

$$\text{Err}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T}\left[\text{tr}\left(\boldsymbol{\Sigma}^2\right) + \left(1 - \frac{1}{T}\right)\left[\text{tr}\left(\boldsymbol{\Sigma}\right)\right]^2\right]. \tag{4.119}$$

The error factors as follows into bias and inefficiency:

$$\text{Inef}^2\left(\widehat{\boldsymbol{\Sigma}}\right) = \frac{1}{T}\left(1 - \frac{1}{T}\right)\left[\text{tr}\left(\boldsymbol{\Sigma}^2\right) + \left[\text{tr}\left(\boldsymbol{\Sigma}\right)\right]^2\right] \tag{4.120}$$

$$\text{Bias}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T^2}\,\text{tr}\left(\boldsymbol{\Sigma}^2\right). \tag{4.121}$$

As expected, the error decreases as the number of observations grows to infinity. Furthermore, it is an increasing function of the average variance: intuitively, more volatile invariants give rise to higher estimation errors. Notice also that the bulk of the error is due to the inefficiency, as the sample estimator is almost unbiased.

From the spectral decomposition (4.112) the following identity follows:

$$\text{tr}\left[\boldsymbol{\Sigma}^2\right] = \text{tr}\left[\boldsymbol{\Lambda}^2\right]. \tag{4.122}$$

Therefore from this expression and (4.114) also the estimation error of the sample covariance (4.119), along with its factorization in terms of bias and inefficiency, is completely determined by the shape of the location-dispersion ellipsoid of the invariants, and not by its location or orientation.

In Figure 4.12 we display the distribution of the loss (4.118) and the respective error (4.119) as the market parameters vary according to (4.116). Notice in the top plot that for high correlations the peak of the distribution of the loss is close to zero, although its dispersion increases dramatically. Indeed, we see in the middle plot how the inefficiency increases with the correlation of the market invariants.

Explicit factors

Consider the particular case of the conditional linear factor model (4.90) where the perturbations are normally distributed:

$$\mathbf{U}_t | \mathbf{f}_t \sim \text{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \tag{4.123}$$

From the expression (2.264) of the density generator:

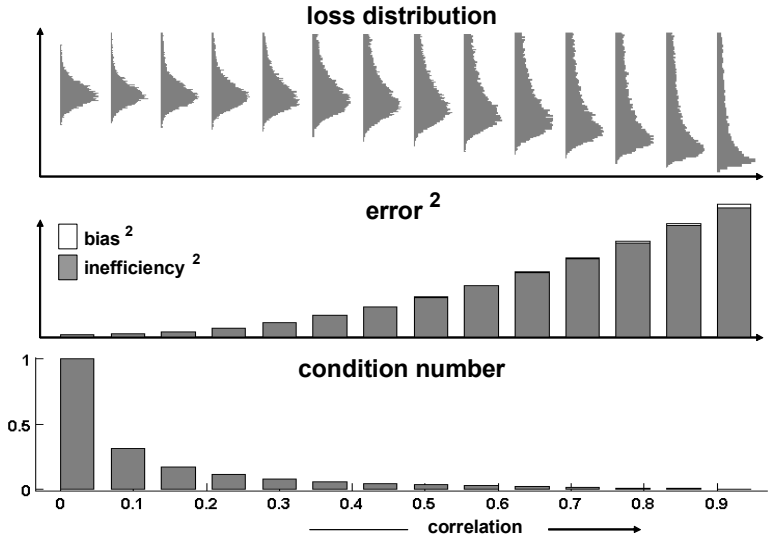

Fig. 4.12. Sample covariance: evaluation

$$g^{\mathrm{N}}(z) \equiv (2\pi)^{-\frac{N}{2}} e^{-\frac{z}{2}}, \tag{4.124}$$

we obtain that the weights (4.94) are constant:

$$w(z) \equiv 1. \tag{4.125}$$

Therefore (4.92) yields the explicit expression of the estimator of the factor loadings in terms of current information:

$$\widehat{\mathbf{B}}[i_T] = \widehat{\mathbf{\Sigma}}_{XF}[i_T]\, \widehat{\mathbf{\Sigma}}_F^{-1}[i_T], \tag{4.126}$$

where

$$\widehat{\mathbf{\Sigma}}_{XF}[i_T] \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \mathbf{f}_t', \quad \widehat{\mathbf{\Sigma}}_F[i_T] \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t'. \tag{4.127}$$

This is the ordinary least squares estimator (4.52) of the regression factor loadings. It is reassuring that two completely different methods yield the same estimator for the factor loadings. This supports our statement that the OLS estimator is the benchmark estimator for the factor loadings.

On the other hand (4.93) yields the explicit expression of the estimator of the dispersion of the perturbations in terms of current information:

$$\widehat{\mathbf{\Sigma}}[i_T] = \frac{1}{T}\sum_{t=1}^{T} \Big(\mathbf{x}_t - \widehat{\mathbf{B}}[i_T]\,\mathbf{f}_t\Big)\Big(\mathbf{x}_t - \widehat{\mathbf{B}}[i_T]\,\mathbf{f}_t\Big)'. \tag{4.128}$$

This is the sample covariance (4.42) of the residuals that stems from the OLS estimation.

To evaluate the goodness of the MLE estimators under the normal hypothesis, we proceed as in (4.15). We prove in Appendix www.4.4 the following results.

The estimator of the factor loadings has a matrix-variate normal distribution:

$$\widehat{\mathbf{B}}\left[I_T|\mathbf{f}_1,\dots,\mathbf{f}_T\right] \sim \mathrm{N}\left(\mathbf{B}, \frac{\boldsymbol{\Sigma}}{T}, \widehat{\boldsymbol{\Sigma}}_F^{-1}\right), \tag{4.129}$$

see (2.181) for the definition of this distribution.

The estimator of the dispersion of the perturbations is a Wishart distributed random matrix (modulo a scale factor):

$$T\widehat{\boldsymbol{\Sigma}}\left[I_T|\mathbf{f}_1,\dots,\mathbf{f}_T\right] \sim \mathrm{W}\left(T-K, \boldsymbol{\Sigma}\right). \tag{4.130}$$

Furthermore, (4.129) and (4.130) are independent of each other.

Given the normal-Wishart joint structure of these estimators, the maximum likelihood estimators of the factor loadings and of the dispersion of the perturbations can be evaluated by exactly the same methodology used for (4.102) and (4.103) respectively.

4.4 Shrinkage estimators

We have discussed in Section 4.2 the benchmark estimators of location and dispersion of the generic market invariants $\mathbf{X}$, namely the sample mean and sample covariance respectively, and the benchmark estimators of the explicit factor models, namely the OLS regression coefficients. These estimators perform well in the limit case of an infinite number of observations, see Figure 4.1. We have also seen in Section 4.3 that when the underlying distribution of the invariants is normal these estimators satisfy the maximum likelihood principle.

Nevertheless, when the sample is very short, the error associated with the benchmark estimators is quite large.

An estimator is *admissible* if it is not systematically outperformed, i.e. if there does not exist another estimator which displays less error for all the stress-test distributions considered in the evaluation of that estimator, see Figure 4.9. The benchmark estimators are not admissible. Indeed, although the maximum likelihood principle is an intuitive recipe with many palatable features, it does not guarantee that the ensuing estimators be optimal.

In particular, the bulk of the error of the benchmark estimators is due to their inefficiency, whereas their bias is quite limited, see Figures 4.11 and 4.12. A key feature of the underlying distribution of the invariants $\mathbf{X}$ that deeply affects the efficiency of the benchmark estimators is the condition number

(4.115), namely the ratio between the smallest and the largest eigenvalues of the unknown underlying scatter matrix:

$$\mathrm{CN}\{\mathbf{X}\} \equiv \frac{\lambda_N}{\lambda_1}. \tag{4.131}$$

We see below that the benchmark estimators are very inefficient when the condition number is close to one, i.e. when the invariants are well-diversified and display little correlation with each other.

In order to fix the inefficiency of the benchmark estimators we consider estimators that are very efficient, although they display a large bias, namely constant estimators. Then we blend the benchmark estimators with the constant estimators by means of weighted averages. Such estimators are called *shrinkage estimators*, because the benchmark estimators are shrunk towards the target constant estimators.

As we see below, the gain in efficiency of the shrinkage estimators with respect to the original benchmark estimators more than compensates for the increase in bias, and thus the overall error of the shrinkage estimators is reduced.

4.4.1 Location

Assume that the market invariants are normally distributed with the following parameters:

$$\mathbf{X}_t \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.132}$$

Consider the standard definition (4.108) of loss of a generic estimator of location $\widehat{\boldsymbol{\mu}}$ with respect to the true unknown location parameter $\boldsymbol{\mu}$ of the invariants:

$$\mathrm{Loss}(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \equiv (\widehat{\boldsymbol{\mu}}[I_T] - \boldsymbol{\mu})' (\widehat{\boldsymbol{\mu}}[I_T] - \boldsymbol{\mu}); \tag{4.133}$$

and the respective definition of error:

$$\mathrm{Err}^2(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \equiv \mathrm{E}\left\{(\widehat{\boldsymbol{\mu}}[I_T] - \boldsymbol{\mu})' (\widehat{\boldsymbol{\mu}}[I_T] - \boldsymbol{\mu})\right\}. \tag{4.134}$$

Consider the benchmark estimator of location (4.98) of the market invariants, namely the sample mean:

$$\widehat{\boldsymbol{\mu}}[i_T] \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t. \tag{4.135}$$

From (4.109) the error (4.134) of the sample mean reads:

$$\mathrm{Err}^2(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \frac{1}{T}\mathrm{tr}(\boldsymbol{\Sigma}). \tag{4.136}$$

In a pathbreaking publication, Stein (1955) proved that the sample mean is not an admissible estimator. In other words, when the dimensions N of the

vector of invariants $\mathbf{X}$ is larger than one, there exists an estimator of location $\widehat{\boldsymbol{\mu}}^S$ such that:

$$\mathrm{Err}^2\left(\widehat{\boldsymbol{\mu}}^S, \boldsymbol{\mu}\right) < \frac{1}{T}\,\mathrm{tr}\left(\boldsymbol{\Sigma}\right), \tag{4.137}$$

no matter the values of the underlying parameters in (4.132). The hypotheses in the original work were somewhat more restrictive than (4.132). Here we discuss the more general case, see also Lehmann and Casella (1998).

First of all, from (4.111) we see that we cannot improve on the sample mean's bias, as the whole error is due to the estimator's inefficiency (4.110). In other words, the sample mean is properly centered around the true, unknown value, but it is too dispersed, see Figure 4.2.

To reduce the error of the estimator we must reduce its inefficiency, although this might cost something in terms of bias. The most efficient estimator is a constant estimator, i.e., an estimator such as (4.12), which with any information associates the same fixed value. Indeed, constant estimators display zero inefficiency, although their bias is very large.

Therefore we consider weighted averages of the sample estimator with a constant estimator of location $\mathbf{b}$, i.e. *any* fixed N-dimensional vector. This way we obtain the *James-Stein shrinkage estimators* of location:

$$\widehat{\boldsymbol{\mu}}^S \equiv (1-\alpha)\,\widehat{\boldsymbol{\mu}} + \alpha\mathbf{b}. \tag{4.138}$$

We show in Appendix www.4.5 that an optimal choice for the weight α in this expression is the following:

$$\alpha \equiv \frac{1}{T}\frac{N\overline{\lambda} - 2\lambda_1}{(\widehat{\boldsymbol{\mu}} - \mathbf{b})'(\widehat{\boldsymbol{\mu}} - \mathbf{b})}, \tag{4.139}$$

where λ_1 is the largest among the N eigenvalues of $\boldsymbol{\Sigma}$ and $\overline{\lambda}$ is the average of the eigenvalues.

By means of *Stein's lemma* we prove in Appendix www.4.5 that the shrinkage estimator (4.138)-(4.139) performs better than the sample mean, i.e. it satisfies (4.137). In real applications the true underlying covariance matrix $\boldsymbol{\Sigma}$ is not known, and thus we cannot compute its eigenvalues. Therefore we replace it with an estimate $\boldsymbol{\Sigma} \mapsto \widehat{\boldsymbol{\Sigma}}$. Furthermore, to obtain more sensible results and to interpret α as a weight, we impose the additional constraint that α be comprised in the interval $(0, 1)$.

As intuition suggests, the optimal amount of shrinkage (4.139) vanishes as the amount of observations T increases.

Furthermore, the optimal shrinkage weight (4.139) is largest in well conditioned market, i.e. when the condition number (4.131) is one. Indeed, in the limit case of full correlation among the invariants, the multivariate setting becomes a one-dimensional setting, in which case the sample estimate is no longer inadmissible.

On the other hand, in the case of extremely well conditioned markets all the eigenvalues are equal to the common value $\overline{\lambda}$ and the optimal shrinkage weight reads:

$$\alpha \equiv \frac{N-2}{T}\frac{\overline{\lambda}}{(\widehat{\boldsymbol{\mu}}-\mathbf{b})'(\widehat{\boldsymbol{\mu}}-\mathbf{b})}. \tag{4.140}$$

Notice in particular that, as intuition suggests, shrinking toward the target $\mathbf{b}$ becomes particularly effective when the number of observations T is low with respect to the dimension of the invariants N and, since $\overline{\lambda}$ is the average variance of the invariants, when the markets are very volatile.

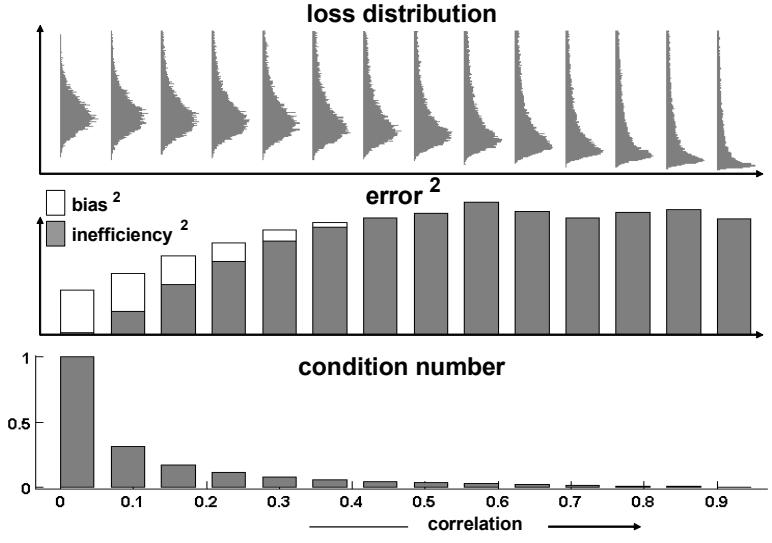

Fig. 4.13. Shrinkage estimator of mean: evaluation

In Figure 4.13 we display the distribution of the loss (4.133) of the shrinkage estimator (4.138)-(4.139) and the respective error (4.134) as the market parameters vary according to (4.116), along with the ensuing condition number. As expected, in well-conditioned markets the amount of shrinkage is maximal. Indeed, the bias is large, whereas the sample mean, which corresponds to a null shrinkage, is unbiased. Nevertheless, the overall error is reduced with respect to the sample mean, compare Figure 4.13 with Figure 4.11.

Shrinking the sample mean towards a constant vector $\mathbf{b}$ is not the only option to improve the estimation. Another possibility consists in shrinking the sample mean towards a scenario-dependent target vector, such as the grand mean. This corresponds to replacing the constant vector $\mathbf{b}$ in (4.138) as follows:

$$\mathbf{b} \mapsto \frac{\mathbf{1}'\widehat{\boldsymbol{\mu}}}{N}\mathbf{1}, \tag{4.141}$$

where $\mathbf{1}$ is an N-dimensional vector of ones.

Another choice of scenario-dependent target is the volatility-weighted grand mean, see Jorion (1986). This corresponds to replacing the constant vector $\mathbf{b}$ in (4.138) as follows:

$$\mathbf{b} \mapsto \frac{\mathbf{1}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}}{\mathbf{1}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}\mathbf{1}. \tag{4.142}$$

where $\widehat{\boldsymbol{\Sigma}}$ is an estimator of the scatter matrix of the invariants.

Several authors have proved the non-admissibility of the sample mean for underlying distributions of the invariants other than (4.132), see Evans and Stark (1996). It is immediate to check that the sample mean is unbiased no matter the underlying distribution, therefore also in the general case an improved estimator must outperform the sample mean in terms of efficiency.

In Chapter 7 we revisit the shrinkage estimators of location in the more general context of Bayesian estimation.

4.4.2 Dispersion and hidden factors

Assume that the market invariants are normally distributed with the following parameters:

$$\mathbf{X}_t \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.143}$$

Consider the standard definition of the loss (4.118) of a generic estimator of dispersion $\widehat{\boldsymbol{\Sigma}}$ with respect to the true unknown underlying scatter parameter $\boldsymbol{\Sigma}$, namely the Frobenius loss:

$$\mathrm{Loss}\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) \equiv \mathrm{tr}\left[\left(\widehat{\boldsymbol{\Sigma}}[I_T] - \boldsymbol{\Sigma}\right)^2\right]; \tag{4.144}$$

and the respective definition of error:

$$\mathrm{Err}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) \equiv \mathrm{E}\left\{\mathrm{tr}\left[\left(\widehat{\boldsymbol{\Sigma}}[I_T] - \boldsymbol{\Sigma}\right)^2\right]\right\}. \tag{4.145}$$

Consider the benchmark estimator of dispersion (4.99), namely the sample covariance:

$$\widehat{\boldsymbol{\Sigma}}[i_T] \equiv \frac{1}{T}\sum_{t=1}^{T}[\mathbf{x}_t - \widehat{\boldsymbol{\mu}}[i_T]][\mathbf{x}_t - \widehat{\boldsymbol{\mu}}[i_T]]', \tag{4.146}$$

where $\widehat{\boldsymbol{\mu}}$ is the sample mean (4.98).

From (4.119) the error (4.145) of the sample covariance reads:

$$\mathrm{Err}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^2\right) + \left(1 - \frac{1}{T}\right)[\mathrm{tr}(\boldsymbol{\Sigma})]^2\right]. \tag{4.147}$$

This is not the minimum error achievable and thus it is possible to define an estimator of dispersion that performs better than the sample covariance.

In order to determine this better estimator we analyze further the error (4.147). Consider as in (4.112) the principal component decomposition of the true unknown scatter matrix:

$$\boldsymbol{\Sigma} \equiv \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'. \tag{4.148}$$

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

$$\boldsymbol{\Lambda} \equiv \text{diag}\,(\lambda_1, \ldots, \lambda_N)\,; \tag{4.149}$$

and the matrix $\mathbf{E}$ is the juxtaposition of the respective orthogonal eigenvectors. Using the identities (4.114) and (4.122), the percentage error (4.31) reads in this context:

$$\text{PErr}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T}\left(1 + \left(1 - \frac{1}{T}\right)\frac{\sum_{n=1}^{N}\lambda_n}{\sum_{n=1}^{N}\lambda_n^2}\right). \tag{4.150}$$

In this expression we can assume without loss of generality that the eigenvalues

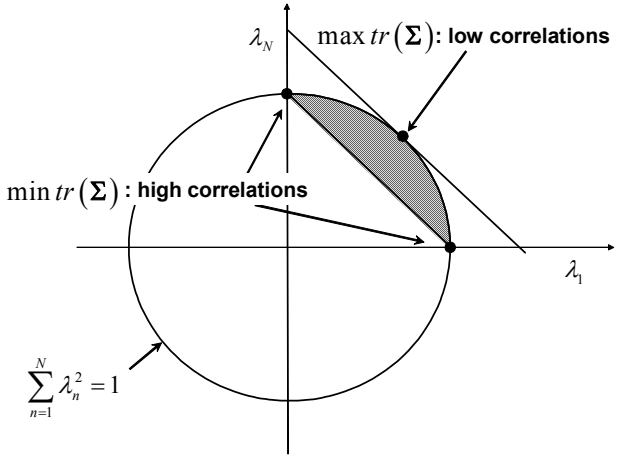

Fig. 4.14. Bounds on the error of the sample covariance matrix

lie on the unit sphere, see Figure 4.14. The sum in the numerator is the trace of $\boldsymbol{\Sigma}$. The different values γ that the trace can assume are represented by the family of hyperplanes (a line in the figure) with equation $\sum_{n=1}^{N}\lambda_n = \gamma$. Since the eigenvalues are constrained on the unit sphere and must be positive, the trace can only span the patterned volume of the hypersphere in Figure 4.14.

The minimum trace corresponds to the following corner solution[5]:

[5] There are actually N solutions, but we only consider one, since we sort the eigenvalues in decreasing order.

$$\lambda_1 = 1, \lambda_2 = \cdots = \lambda_N = 0 \Leftrightarrow \mathrm{tr}(\boldsymbol{\Sigma}) = 1, \tag{4.151}$$

which gives rise to a condition number (4.131) equal to zero. In this situation the ellipsoid $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ determined by the market parameters as described in $(A.73)$, which is also the location-dispersion ellipsoid of the invariants (2.75), is squeezed into a line. In other words, there exists only one actual dimension of risk, as all the invariants can be expressed as functions of one specific invariant. This is approximately the case in the swap market, as we see in Figure 3.20. In this environment of high correlations the percentage estimation error is minimal, and reads:

$$\mathrm{PErr}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T}\left(2 - \frac{1}{T}\right). \tag{4.152}$$

On the other hand, the maximum trace corresponds to the following combination:

$$\lambda_1 = \cdots = \lambda_N = \frac{1}{\sqrt{N}} \Leftrightarrow \mathrm{tr}(\boldsymbol{\Sigma}) = \sqrt{N}, \tag{4.153}$$

which gives rise to a condition number (4.131) equal to one. In this case the location-dispersion ellipsoid of the market invariants becomes a sphere, which means that cross-correlations among the invariants are zero. Therefore, in a zero-correlation environment, the percentage error is maximal and reads:

$$\mathrm{PErr}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right) = \frac{1}{T}\left(1 + \left(1 - \frac{1}{T}\right)N\right). \tag{4.154}$$

Notice that the estimation degenerates as the dimension N of the invariants becomes large as compared with the number T of observations.

To summarize, we need an estimator that improves on the sample covariance especially when the market invariants are well conditioned and when the number of observations in the sample is small with respect to the number of invariants.

To introduce this estimator, we notice from (4.120)-(4.121) that the sample covariance's bias is minimal, as almost the whole error is due to the estimator's inefficiency. In other words, the sample covariance is properly centered around the true, unknown value, but it is too dispersed, see Figure 4.2.

The sample covariance is inefficient because the estimation process tends to scatter the sample eigenvalues $\widehat{\boldsymbol{\Lambda}}$ away from the mean value $\overline{\lambda}$ of the true unknown eigenvalues. Indeed, Ledoit and Wolf (2004) prove the following general result:

$$\mathrm{E}\left\{\sum_{n=1}^{N}\left(\widehat{\lambda}_n - \overline{\lambda}\right)^2\right\} = \sum_{n=1}^{N}\left(\lambda_n - \overline{\lambda}\right)^2 + \mathrm{Err}^2\left(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}\right). \tag{4.155}$$

Geometrically, the estimation process squeezes and stretches the location-dispersion ellipsoid $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ of the market invariants.

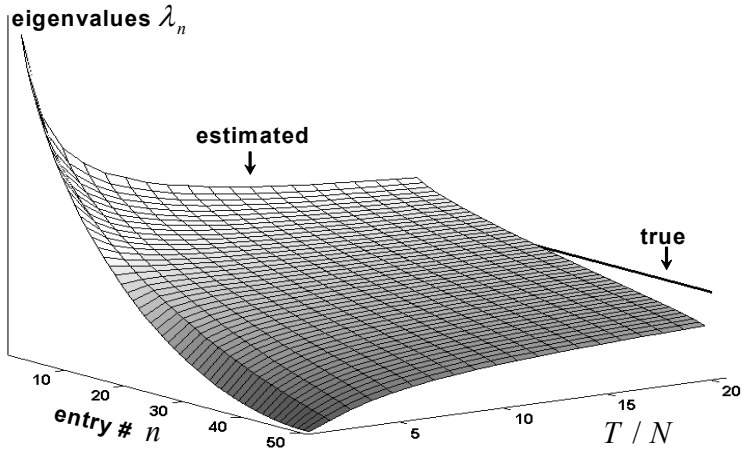

Fig. 4.15. Scattering of sample eigenvalues

Since the estimation error is large when the number of observations T is small, the scattering effect is larger when T is small with respect to the number of invariants. We plot this phenomenon in Figure 4.15 for the case of $N \equiv 50$ market invariants. As we show in Appendix www.4.6, in the extreme case where the number of observations T is lower than the number of invariants N, the last sample eigenvalues become null and thus the sample covariance becomes singular.

Furthermore, the scattering of the eigenvalues of the sample covariance is more pronounced for those invariants whose location-dispersion ellipsoid is close to a sphere. This result is intuitive: comparatively speaking, a sphere gets squeezed and stretched more than an elongated ellipsoid, which is already elongated to begin with. This result is also consistent with (4.152) and (4.154).

We can summarize the cause of the inefficiency of the sample covariance in terms of the condition number (4.131). Indeed, the estimation process worsens the condition number of the market invariants:

$$\widehat{\mathrm{CN}}\{\mathbf{X}\} \equiv \frac{\widehat{\lambda}_N}{\widehat{\lambda}_1} < \frac{\lambda_N}{\lambda_1} \equiv \mathrm{CN}\{\mathbf{X}\}. \tag{4.156}$$

To reduce the error of the sample covariance we must reduce its inefficiency by averaging it with an efficient and well conditioned estimator of dispersion. On the one hand, the most efficient estimator is a constant estimator, i.e. an estimator such as (4.12), which with any information associates a given fixed value: indeed, constant estimators display zero inefficiency, although their bias is very large. On the other hand, the best-conditioned matrices are multiples of the identity, in which case the condition number is one.

Therefore, the ideal candidate to reduce the inefficiency of the sample covariance is the following constant, well conditioned matrix:

$$\mathbf{C} \equiv \overline{\lambda}\mathbf{I}_N, \tag{4.157}$$

where the mean value $\overline{\lambda}$ of the true unknown eigenvalues represents the average variance of the invariants:

$$\overline{\lambda} \equiv \frac{\mathrm{tr}\{\boldsymbol{\Lambda}\}}{N} = \frac{\mathrm{tr}\{\boldsymbol{\Sigma}\}}{N} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Var}\{X_n\}. \tag{4.158}$$

Nevertheless, the true eigenvalues are unknown, therefore we replace (4.157) with its sample counterpart:

$$\widehat{\mathbf{C}} \equiv \frac{\sum_{n=1}^{N} \widehat{\lambda}_n}{N}\mathbf{I}. \tag{4.159}$$

At this point, following Ledoit and Wolf (2004) we define the *shrinkage estimator of dispersion* as the weighted average of the sample covariance and the target matrix:

$$\widehat{\boldsymbol{\Sigma}}^S \equiv (1-\alpha)\,\widehat{\boldsymbol{\Sigma}} + \alpha\widehat{\mathbf{C}}. \tag{4.160}$$

The optimal shrinkage weight in this expression is defined as follows:

$$\alpha \equiv \frac{1}{T}\frac{\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\left\{\left(\mathbf{x}_t\mathbf{x}_t' - \widehat{\boldsymbol{\Sigma}}\right)^2\right\}}{\mathrm{tr}\left\{\left(\widehat{\boldsymbol{\Sigma}} - \widehat{\mathbf{C}}\right)^2\right\}}, \tag{4.161}$$

if $\alpha < 1$, and 1 otherwise.

The shrinkage estimator (4.160) is indeed better conditioned than the sample covariance:

$$\frac{\widehat{\lambda}_N^S}{\widehat{\lambda}_1^S} > \frac{\widehat{\lambda}_N}{\widehat{\lambda}_1}, \tag{4.162}$$

see Appendix www.4.6. Thus the ensuing error (4.145) is less than for the sample covariance.

As intuition suggests, the optimal amount of shrinkage (4.161) vanishes as the amount of observations T increases.

Furthermore, the optimal shrinkage weight is largest when the condition number of the market invariants is close to one. Indeed, in this case the denominator in (4.161) becomes very small. This is consistent with the fact that the percentage error is maximal in well-condition markets, see (4.154).

In Figure 4.16 we display the distribution of the loss (4.144) of the shrinkage estimator (4.160) and the respective error (4.145) as the market parameters vary according to (4.116), along with the ensuing condition number. Notice that shrinking towards a multiple of the identity matrix introduces a

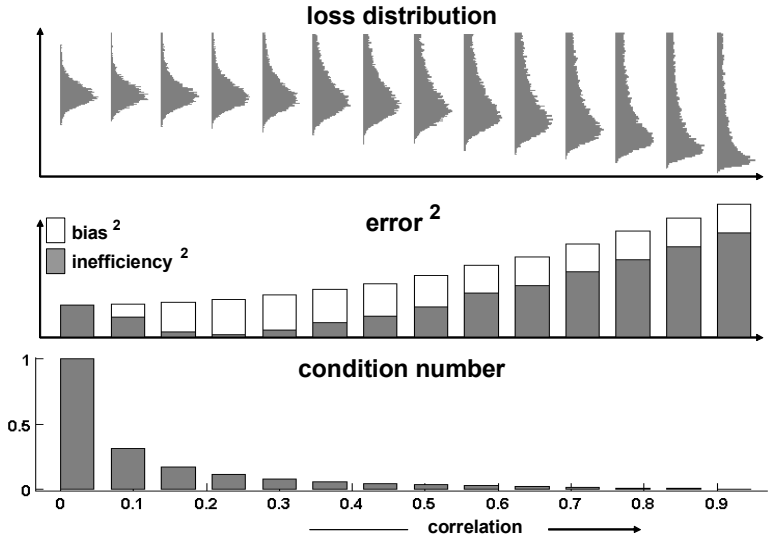

Fig. 4.16. Shrinkage estimator of covariance: evaluation

bias that was not present in the case of the sample covariance, see Figure 4.12. Nevertheless, the overall error is reduced by the shrinkage process.

In Chapter 7 we revisit the shrinkage estimators of dispersion in the more general context of Bayesian estimation.

4.4.3 Explicit factors

The benchmark estimator of the factor loadings in an explicit factor model is the ordinary least square estimator of the regression coefficients (4.126) and the estimator of the dispersion of the residuals is the respective sample covariance matrix (4.128). Like in the case of the estimators of location and dispersion, it is possible to improve on these estimators by shrinking them towards suitable targets, see Ledoit and Wolf (2003) for an application of a one-factor model to the stock market.

We discuss in Chapter 7 the shrinkage estimators of explicit-factor models in the more general context of Bayesian estimation.

4.5 Robustness

In our journey throughout the possible approaches to building estimators we have always assumed that the true, unknown distribution of the market invariants lies somewhere in the subset of stress test distributions, refer to Figure

4.3 for the general case and to Figure 4.9 for the parametric approach. In this section we discuss *robust estimation*, which deals with the potential consequences and possible remedies of choosing a space of stress test distributions that does not include the true, unknown distribution of the market invariants, see Figure 4.17.

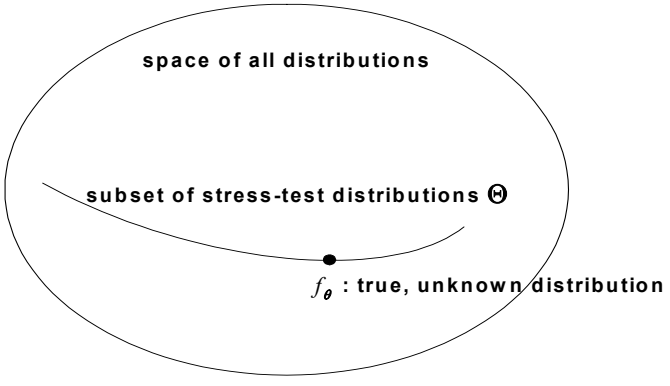

Fig. 4.17. Robust approach to estimation

To provide the intuition behind robust estimation, consider as in Figure 4.7 the location-dispersion ellipsoid (2.75) defined by the sample-mean (4.41) and sample-covariance (4.42) of a set of observations of market invariants:

$$\mathcal{E}_{\widehat{\mathrm{E}},\widehat{\mathrm{Cov}}} \equiv \left\{ \mathbf{x} \in \mathbb{R}^N \text{ such that } \left(\mathbf{x} - \widehat{\mathrm{E}}\right)' \left(\widehat{\mathrm{Cov}}\right)^{-1} \left(\mathbf{x} - \widehat{\mathrm{E}}\right) \leq 1 \right\}. \quad (4.163)$$

Then add a fake observation, an outlier, and repeat the estimation based on the enlarged sample. The new ellipsoid, which represents the new sample mean and sample covariance, is completely different, see Figure 4.18: one single observation completely disrupted the estimation.

In the above experiment we know that the extra-observation is spurious. Therefore, such an extreme sensitivity does not represent a problem. If we knew for a fact that some observations were spurious, a sensitive estimator would help us detect the unwelcome outliers. This is the subject of outlier detection, which we tackle in Section 4.6.1.

On the other hand, in many applications we do not know the true underlying distribution and, most importantly, we have no reason to believe that some observations could be spurious. Therefore we cannot trust sensitive estimators such as the sample estimators. Instead, we need to develop estimators that properly balance the trade-off between the precision and the robustness of the final results.

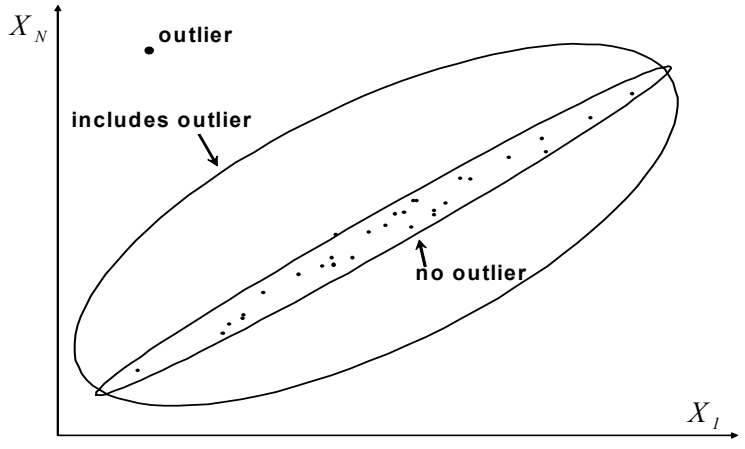

Fig. 4.18. Sample estimators: lack of robustness

In this section, first we discuss a few measures of robustness for an estimator, namely the jackknife, the sensitivity curve and, most notably, the influence function: when this is bounded, the respective estimator is robust.

Then we compute the influence function of the estimators introduced so far, namely nonparametric sample estimators and parametric maximum likelihood estimators of location, dispersion and explicit factor loadings. As it turns out, the sample estimators display an unbounded influence function and therefore they are not robust. On the other hand, the maximum likelihood estimators display a range of behaviors: for instance, MLE of normally distributed invariants are the sample estimators and therefore they are not robust. On the other hand, MLE of Cauchy-distributed invariants have bounded influence function and therefore they are robust.

Finally, we show how to build robust estimators of the main parameters of interest for asset allocation problem.

4.5.1 Measures of robustness

To tackle robustness issues, we need first of all to be able to measure the robustness of a generic estimator. First we introduce two qualitative measures, namely the jackknife and the sensitivity curve. Relying on the intuition behind these measures we introduce a tool that precisely quantifies the robustness of an estimator, namely the influence function.

Consider a generic estimator $\widehat{\mathbf{G}}$ of S features of an unknown distribution. As in (4.9), an estimator is a vector-valued function of currently available information, which is represented as in (4.8) by the time series of the past occurrences of the market invariants:

$$i_T \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \mapsto \widehat{\mathbf{G}}. \tag{4.164}$$

A first measure of robustness of an estimator is the *jackknife*, introduced by Quenouille (1956) and Tukey (1958). The jackknife is built as follows. First we remove the generic t-th observation from the time series; then we estimate the quantity of interest from the reduced time series:

$$\widehat{\mathbf{G}}_{(-t)} \equiv \widehat{\mathbf{G}}\left(\mathbf{x}_1 \ldots, \mathbf{x}_{t-1}, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_T\right); \tag{4.165}$$

finally we put back in place the t-th observation. We repeat this process for all the observations, computing a total of T estimates. If all the estimates are comparable, we assess that the estimator is robust. In Figure 4.18 we see that this is not the case for the sample mean and the sample covariance.

To build another measure of robustness, instead of removing in turn all the observations, we can add an arbitrary observation to the time series and evaluate its effect on the estimate. This way we obtain the *sensitivity curve*, introduced by Tukey (1977) and defined as follows:

$$\mathrm{SC}\left(\mathbf{x}, \widehat{\mathbf{G}}\right) \equiv T\widehat{\mathbf{G}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{x}\right) - T\widehat{\mathbf{G}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_T\right), \tag{4.166}$$

where the normalization T is meant to make the evaluation less sensitive to the sample size. If the sensitivity curve is small for any value of the extra-observation $\mathbf{x}$, we assess that the estimator $\widehat{\mathbf{G}}$ is robust. We see in Figure 4.18 that this is not the case for the sample mean and the sample covariance.

Both jackknife and sensitivity curve are qualitative tools that can detect lack of robustness: if either measure shows that the given estimator is not robust, we should reject that estimator and search for a better one.

Nevertheless, if an estimator is not rejected, we cannot draw any conclusion on the degree of robustness of that estimator. Indeed, as far as the sensitivity curve is concerned, whatever result we obtain depends on the specific sample. On the other hand, as far as the jackknife is concerned, the sample might contain two or more outliers instead of one: in this case we might consider tests that remove more than one observation at a time, but we would not be sure where to stop.

To obtain a tool that quantifies robustness independently of the specific sample, we should move in the opposite direction, considering the marginal effect of an outlier when the sample size tends to infinity. The *influence function* can be defined heuristically as the infinite-sample limit of the sensitivity curve, see Hampel, Ronchetti, Rousseeuw, and Stahel (1986). Intuitively, the influence function quantifies the marginal effect on an estimator of an extra-observation in the limit of infinite observations.

In order to introduce this limit, we need to express the generic S-dimensional estimator as an S-dimensional functional of the empirical probability density function:

$$\widehat{\mathbf{G}} \equiv \widetilde{\mathbf{G}}\left[f_{i_T}\right], \tag{4.167}$$

where the empirical probability density function (2.240) is defined in terms of the Dirac delta ($B.16$) as follows:

$$f_{i_T} \equiv \frac{1}{T}\sum_{t=1}^{T} \delta^{(\mathbf{x}_t)}. \tag{4.168}$$

The sample estimators are explicit functionals of the empirical probability density function. Indeed the sample estimators aim at estimating some functional $\mathbf{G}[f_{\mathbf{X}}]$ of the unknown probability density function $f_{\mathbf{X}}$ of the market invariants. Therefore by their very definition (4.36) the functional that defines the estimator is the functional that defines the quantity of interest of the unknown distribution of the market invariants:

$$\widehat{\mathbf{G}} \equiv \mathbf{G}[f_{i_T}]. \tag{4.169}$$

This expression is clearly in the form (4.167).

For example, consider the following functional:

$$\mathbf{G}[h] \equiv \int_{\mathbb{R}^N} \mathbf{x} h(\mathbf{x})\, d\mathbf{x}, \tag{4.170}$$

where h is any function such that the integral (4.170) makes sense. This functional, when applied to the probability density function of a distribution, yields its expected value:

$$\mathbf{G}[f_{\mathbf{X}}] = \mathrm{E}\{\mathbf{X}\}. \tag{4.171}$$

Consider now the sample mean:

$$\widehat{\mathbf{G}} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t. \tag{4.172}$$

The sample mean is the functional (4.170) applied to the empirical pdf:

$$\widehat{\mathbf{G}} = \int_{\mathbb{R}^N} \mathbf{x} f_{i_T}(\mathbf{x})\, d\mathbf{x} \equiv \mathbf{G}[f_{i_T}]. \tag{4.173}$$

On the other hand, the maximum likelihood estimators are implicit functionals of the empirical probability density function. Consider the ML estimator $\widehat{\boldsymbol{\theta}}$ of the S-dimensional parameter $\boldsymbol{\theta}$ of a distribution $f_{\boldsymbol{\theta}}$. The ML estimator as a functional is defined implicitly by the first-order conditions on the log-likelihood. Indeed, from their definition (4.66) the ML estimators solve in quite general cases the following implicit equation:

$$\mathbf{0} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\psi}\left(\mathbf{x}_t, \widehat{\boldsymbol{\theta}}\right), \tag{4.174}$$

where $\boldsymbol{\psi}$ is the S-dimensional vector of first-order partial derivatives of the log-likelihood:

$$\boldsymbol{\psi}(\mathbf{x},\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} \ln\left(f_{\boldsymbol{\theta}}(\mathbf{x})\right). \tag{4.175}$$

Consider now the S-dimensional functional $\widetilde{\boldsymbol{\theta}}[h]$ defined implicitly for a generic function h in a suitable domain as follows:

$$\widetilde{\boldsymbol{\theta}}[h]: \quad \int_{\mathbb{R}^N} \boldsymbol{\psi}\left(\mathbf{x},\widetilde{\boldsymbol{\theta}}\right) h(\mathbf{x})\, d\mathbf{x} \equiv \mathbf{0}. \tag{4.176}$$

In this notation, the ML estimator (4.174) can be written as follows:

$$\widehat{\boldsymbol{\theta}} \equiv \widetilde{\boldsymbol{\theta}}[f_{i_T}], \tag{4.177}$$

which is in the form (4.167).

For example, consider the functional $\widetilde{\theta}[h]$ defined implicitly by the following equation:

$$\widetilde{\theta}[h]: \quad 0 \equiv \int_{\mathbb{R}} \left(\ln x - \widetilde{\theta}\right) h dx. \tag{4.178}$$

Now assume as in (4.57) that there exists a lognormally distributed invariant with the following parameters:

$$X \sim \mathrm{LogN}(\theta, 1). \tag{4.179}$$

The ML estimator of θ reads:

$$\widehat{\theta} = \frac{1}{T}\sum_{t=1}^{T} \ln x_t, \tag{4.180}$$

see (4.69). Clearly, the ML estimator of θ solves:

$$\begin{aligned} 0 &= \frac{1}{T}\sum_{t=1}^{T}\left(\ln x_t - \widehat{\theta}\right) \\ &= \int_{\mathbb{R}} \left(\ln x - \widehat{\theta}\right) \frac{1}{T}\sum_{t=1}^{T} \delta^{(x_t)}(x)\, dx \\ &= \int_{\mathbb{R}} \left(\ln x - \widehat{\theta}\right) f_{i_T}(x)\, dx. \end{aligned} \tag{4.181}$$

Therefore:

$$\widehat{\theta} = \widetilde{\theta}[f_{i_T}]. \tag{4.182}$$

Notice that the term in brackets in the integral (4.178) is the first-order derivative of the logarithm of the probability density function (4.58), as prescribed by (4.175).

Consider the sensitivity curve (4.166). Adding one observation in an arbitrary position $\mathbf{x}$ corresponds to modifying the empirical probability density function (4.168) as follows:

$$f_{i_T} \mapsto (1-\epsilon) f_{i_T} + \epsilon \delta^{(\mathbf{x})}, \tag{4.183}$$

where $\epsilon \equiv 1/(T+1)$ is the relative weight of the extra-observation and δ is the Dirac delta (B.16). Therefore in the functional notation (4.167) the sensitivity curve (4.166) reads:

$$\mathrm{SC}\left(\mathbf{x}, \widehat{\mathbf{G}}\right) \equiv \frac{1-\epsilon}{\epsilon}\left\{\widetilde{\mathbf{G}}\left[(1-\epsilon) f_{i_T} + \epsilon \delta^{(\mathbf{x})}\right] - \widetilde{\mathbf{G}}\left[f_{i_T}\right]\right\}. \tag{4.184}$$

In the limit of an infinite number of observations T the relative weight ϵ tends to zero. Furthermore, from the Glivenko-Cantelli theorem (4.34) the empirical pdf f_{i_T} tends to the true, unknown pdf $f_{\mathbf{X}}$. Therefore the influence function of a generic S-dimensional estimator $\widehat{\mathbf{G}}$, which is the infinite-sample limit of the sensitivity curve, is defined as the following S-dimensional vector:

$$\mathrm{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\mathbf{G}}\right) \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\widetilde{\mathbf{G}}\left[(1-\epsilon) f_{\mathbf{X}} + \epsilon \delta^{(\mathbf{x})}\right] - \widetilde{\mathbf{G}}\left[f_{\mathbf{X}}\right]\right), \tag{4.185}$$

where $\widetilde{\mathbf{G}}$ is the S-dimensional functional (4.167) that links the estimator to the empirical probability density function.

In order to use the influence function in applications, we need to define it more formally as a Gateaux derivative, which is the equivalent of a partial derivative in the world of functional analysis.

We recall that the partial derivatives of a function g defined in $\mathbb{R}^N$ at the point $\mathbf{v}$ are N numbers D, commonly denoted as follows:

$$D(n, \mathbf{v}, g) \equiv \frac{\partial g(\mathbf{v})}{\partial v_n}, \quad n = 1, \ldots, N. \tag{4.186}$$

These N numbers are such that such that whenever $\mathbf{u} \approx \mathbf{v}$ the following approximation holds:

$$g(\mathbf{u}) - g(\mathbf{v}) \approx \sum_{n=1}^{N} D(n, \mathbf{v}, g)(u_n - v_n). \tag{4.187}$$

According to Table B.4, in the world of functional analysis the vector's index n is replaced by the function's argument $\mathbf{x}$, vectors such as $\mathbf{v}$ are replaced by functions $v(\cdot)$ and sums are replaced by integrals. Furthermore, functions g are replaced with functionals G.

The *Gateaux derivative* is the partial derivative (4.187) in this new notation. In other words it is the number D such that whenever two functions are close $u \approx v$ the following approximation holds:

$$G[u] - G[v] \approx \int_{\mathbb{R}^N} D(\mathbf{x}, v, G)\, u(\mathbf{x})\, d\mathbf{x}, \tag{4.188}$$

where we used the normalization:

$$\int_{\mathbb{R}^N} D(\mathbf{x}, v, G)\, v d\mathbf{x} \equiv 0. \tag{4.189}$$

Consider an estimator $\widehat{\mathbf{G}}$ that is represented by the functional $\widetilde{\mathbf{G}}$ as in (4.167). The influence function of each entry of the estimator $\widehat{\mathbf{G}}$ for a given distribution $f_{\mathbf{X}}$ is the Gateaux derivative of the respective entry of $\widetilde{\mathbf{G}}$ in $f_{\mathbf{X}}$:

$$\mathrm{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\mathbf{G}}\right) \equiv \mathbf{D}\left(\mathbf{x}, f_{\mathbf{X}}, \widetilde{\mathbf{G}}\right). \tag{4.190}$$

Indeed, setting $u \equiv (1-\epsilon) f_{\mathbf{X}} + \epsilon \delta^{(\mathbf{x})}$ in (4.188) yields the heuristic definition (4.185).

An estimator is robust if its influence function is small, or at least bounded, as the extra observation $\mathbf{x}$ varies in a wide range in the space of observations and as the distribution of the invariants $f_{\mathbf{X}}$ varies in a wide range in the space of distributions.

More precisely, suppose that we are interested in some parameters $\mathbf{G}[f_{\mathbf{X}}]$ of the unknown distribution $f_{\mathbf{X}}$ of the market invariants. As usual, we make assumptions on the set of possible distributions for $f_{\mathbf{X}}$ and we build an estimator $\widehat{\mathbf{G}}$. Suppose that we choose inappropriately a family of stress test distributions that does not include the true, unknown distribution $f_{\mathbf{X}}$, i.e. we miss the target by some extent as in Figure 4.17. Under these "wrong" assumptions we develop the "wrong" estimator $\widehat{\mathbf{G}}$, which can be expressed as a functional of the empirical pdf as in (4.167). The influence function provides a measure of the damage:

$$\widehat{\mathbf{G}} - \mathbf{G}[f_{\mathbf{X}}] \approx \frac{1}{T}\sum_{t=1}^{T} \mathrm{IF}\left(\mathbf{x}_t, f_{\mathbf{X}}, \widehat{\mathbf{G}}\right), \tag{4.191}$$

where the approximation improves with the number of observations. This follows immediately by setting $u \equiv f_{i_T}$ in (4.188) and using the fact that estimators are typically *Fisher consistent*, i.e. such that:

$$\widetilde{\mathbf{G}}[f_{\mathbf{X}}] = \mathbf{G}[f_{\mathbf{X}}]. \tag{4.192}$$

Of course, we do not know the true underlying distribution $f_{\mathbf{X}}$ of the market invariants, but as long as the influence function is bounded for a wide range of underlying distributions $f_{\mathbf{X}}$, the damage is contained.

4.5.2 Robustness of previously introduced estimators

In Section 4.2 we introduced the nonparametric sample estimators $\widehat{\mathbf{G}}$ of the unknown features $\mathbf{G}[f_{\mathbf{X}}]$ of the distribution of the market invariants. The functional representation $\widetilde{\mathbf{G}}[f_{i_T}]$ of sample estimators in terms of the empirical probability density function is explicit and defined by (4.169). Therefore,

the expression of the influence function of generic nonparametric estimators follows directly from the heuristic definition (4.185) of the influence function and reads:

$$\text{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\mathbf{G}}\right) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\mathbf{G}\left[(1-\epsilon) f_{\mathbf{X}} + \epsilon \delta^{(\mathbf{x})}\right] - \mathbf{G}\left[f_{\mathbf{X}}\right]\right). \tag{4.193}$$

In Section 4.3 we introduced the maximum likelihood estimators of the parameters $\boldsymbol{\theta}$ of the distribution $f_{\boldsymbol{\theta}}$ of the market invariants. The functional representation $\widetilde{\boldsymbol{\theta}}\left[f_{i_T}\right]$ of the maximum likelihood estimators in terms of the empirical probability density function is implicit and defined by (4.176). We prove in Appendix www.4.7 that in this case the influence function reads:

$$\text{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\boldsymbol{\theta}}\right) = \mathbf{A}\left.\frac{\partial \ln f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} \equiv \widetilde{\boldsymbol{\theta}}\left[f_{\mathbf{X}}\right]}, \tag{4.194}$$

where the constant $S \times S$ matrix $\mathbf{A}$ is defined as follows:

$$\mathbf{A} \equiv -\left[\int_{\mathbb{R}^N} \left.\frac{\partial^2 \ln f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} \equiv \widetilde{\boldsymbol{\theta}}\left[f_{\mathbf{X}}\right]} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\right]^{-1}. \tag{4.195}$$

We proceed below to apply these formulas to the sample and maximum likelihood estimators of interest for asset allocation problems.

Location and dispersion

Consider the sample estimators of location and dispersion of the market invariants $\mathbf{X}_t$, i.e. the sample mean (4.41) and the sample covariance (4.42) respectively:

$$\widehat{\mathrm{E}} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t \tag{4.196}$$

$$\widehat{\mathrm{Cov}} \equiv \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}_t - \widehat{\mathrm{E}}\right)\left(\mathbf{x}_t - \widehat{\mathrm{E}}\right)'. \tag{4.197}$$

We prove in Appendix www.4.7 that the influence function (4.193) for the sample mean reads:

$$\text{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\mathrm{E}}\right) = \mathbf{x} - \mathrm{E}\{\mathbf{X}\}; \tag{4.198}$$

and the influence function (4.193) for the sample covariance reads:

$$\text{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\mathrm{Cov}}\right) = (\mathbf{x} - \mathrm{E}\{\mathbf{X}\})(\mathbf{x} - \mathrm{E}\{\mathbf{X}\})' - \mathrm{Cov}\{\mathbf{X}\}. \tag{4.199}$$

Notice that the influence function of the sample estimators is not bounded. Therefore, the sample estimators are not robust: a strategically placed outlier,

also known as *leverage point*, can completely distort the estimation. This is the situation depicted in Figure 4.18.

Assume now that the invariants $\mathbf{X}_t$ are elliptically distributed:

$$\mathbf{X}_t \sim \mathrm{El}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g\right), \tag{4.200}$$

where $\boldsymbol{\mu}$ is the N-dimensional location parameter, $\boldsymbol{\Sigma}$ is the $N \times N$ dispersion matrix and g is the probability density generator. In other words, the probability density of the invariants $\mathbf{X}$ reads:

$$f_{\boldsymbol{\theta}}\left(\mathbf{x}\right) \equiv \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} g\left(\mathrm{Ma}^{\,2}\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right), \tag{4.201}$$

where $\mathrm{Ma}\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the Mahalanobis distance of the point $\mathbf{x}$ from the point $\boldsymbol{\mu}$ through the metric $\boldsymbol{\Sigma}$:

$$\mathrm{Ma}^2\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \left(\mathbf{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}-\boldsymbol{\mu}\right). \tag{4.202}$$

In this case the parametric distribution of the market invariants is fully determined by the set of parameters is $\boldsymbol{\theta} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Consider the maximum likelihood estimators of the parameters $\boldsymbol{\theta}$, which are defined by the implicit equations (4.77)-(4.79) as follows:

$$\widehat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \frac{w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)}{\sum_{s=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_s, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)} \mathbf{x}_t \tag{4.203}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}\right)' w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right), \tag{4.204}$$

where

$$w\left(z\right) \equiv -2\frac{g'\left(z\right)}{g\left(z\right)}. \tag{4.205}$$

These parameters can be expressed as functionals $\widetilde{\boldsymbol{\mu}}\left[f_{i_T}\right]$ and $\widetilde{\boldsymbol{\Sigma}}\left[f_{i_T}\right]$ of the empirical pdf. The functionals are defined implicitly as in (4.176) as follows:

$$\int_{\mathbb{R}^N} \boldsymbol{\psi}\left(\mathbf{x}, \widetilde{\boldsymbol{\mu}}\left[h\right], \widetilde{\boldsymbol{\Sigma}}\left[h\right]\right) h\left(\mathbf{x}\right) d\mathbf{x} \equiv \mathbf{0}. \tag{4.206}$$

The vector-valued function $\boldsymbol{\psi}$ in this expression follows from (4.203)-(4.204) and reads:

$$\boldsymbol{\psi}\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \begin{pmatrix} w\left(\mathrm{Ma}^2\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right)\left(\mathbf{x}-\boldsymbol{\mu}\right) \\ w\left(\mathrm{Ma}^2\left(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \mathrm{vec}\left[\left(\left(\mathbf{x}-\boldsymbol{\mu}\right)\left(\mathbf{x}-\boldsymbol{\mu}\right)' - \boldsymbol{\Sigma}\right)\right] \end{pmatrix}, \tag{4.207}$$

where vec is the operator $(A.104)$ that stacks the columns of a matrix into a vector. From (4.194) and (4.175) the norm of the influence function is proportional to the norm of the above vector:

$$\left\| \text{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)\right\| \propto \|\boldsymbol{\psi}\|. \tag{4.208}$$

In particular, if the invariants are normally distributed the term w in (4.207) becomes $w \equiv 1$, see (4.97). Therefore the influence function is not bounded. This is not surprising, since we know from Section 4.3 that the ML estimators of location and dispersion of normally distributed invariants are the sample estimators and thus their influence function is (4.198)-(4.199). In other words, the ML estimators of location and dispersion of normally distributed invariants are not robust.

On the other hand, if the invariants are elliptically but not normally distributed the influence function displays a different behavior. Consider for example Cauchy-distributed invariants. In this case from (4.84) the term w in (4.207) becomes:

$$w(z) = \frac{N+1}{1+z}. \tag{4.209}$$

Therefore, from (4.208) and (4.202) the influence function of the location and dispersion maximum likelihood estimators becomes bounded. In other words, the ML estimators of location and dispersion of Cauchy-distributed invariants are robust.

Explicit factors

Consider an explicit factor linear model:

$$\mathbf{X}_t = \mathbf{B}\mathbf{F}_t + \mathbf{U}_t. \tag{4.210}$$

The sample estimator of the regression factor loadings are the ordinary least squares coefficients (4.52), which we report here:

$$\widehat{\mathbf{B}} \equiv \left(\sum_t \mathbf{x}_t \mathbf{f}_t'\right)\left(\sum_t \mathbf{f}_t \mathbf{f}_t'\right)^{-1}. \tag{4.211}$$

We do not discuss the sample covariance of the perturbation, which is the same as (4.197), where $\widehat{\mathbf{B}}\mathbf{f}_t$ replaces $\widehat{\mathrm{E}}$. We prove in Appendix www.4.7 that the influence function for the OLS coefficients reads:

$$\text{IF}\left((\mathbf{x}, \mathbf{f}), f_{\mathbf{X},\mathbf{F}}, \widehat{\mathbf{B}}\right) = \left(\mathbf{x}\mathbf{f}' - \mathbf{B}\mathbf{f}\mathbf{f}'\right) \mathrm{E}\left\{\mathbf{F}\mathbf{F}'\right\}^{-1}. \tag{4.212}$$

Notice that the influence function of the sample OLS coefficients is not bounded. Therefore, the OLS estimate is not robust: a strategically placed outlier, also known as *leverage point*, can completely distort the estimation. This is the situation depicted in Figure 4.18.

Consider now a parametric explicit factor model conditioned on the factors. We assume as in (4.90) that the perturbations are elliptically distributed

and centered in zero. Therefore the respective conditional explicit factor model reads:

$$\mathbf{X}_t|\mathbf{f}_t \sim \mathrm{El}\left(\mathbf{B}\mathbf{f}_t, \boldsymbol{\Sigma}, g\right), \tag{4.213}$$

where $\boldsymbol{\Sigma}$ is the $N \times N$ dispersion matrix of the perturbations and g is their probability density generator. In this case the parametric distribution of the market invariants is fully determined by the set of parameters is $\boldsymbol{\theta} \equiv (\mathbf{B}, \boldsymbol{\Sigma})$.

Consider the maximum likelihood estimators of the parameters $\boldsymbol{\theta}$, which are defined by the implicit equations (4.92)-(4.94) as follows:

$$\begin{aligned}\widehat{\mathbf{B}} = &\left[\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\mathbf{x}_t\mathbf{f}_t'\right] \\ &\left[\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\mathbf{f}_t\mathbf{f}_t'\right]^{-1}\end{aligned} \tag{4.214}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T} w\left(\mathrm{Ma}^2\left(\mathbf{x}_t, \widehat{\mathbf{B}}\mathbf{f}_t, \widehat{\boldsymbol{\Sigma}}\right)\right)\left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)\left(\mathbf{x}_t - \widehat{\mathbf{B}}\mathbf{f}_t\right)'. \tag{4.215}$$

where

$$w(z) \equiv -2\frac{g'(z)}{g(z)}. \tag{4.216}$$

These parameters can be expressed as functionals $\widetilde{\mathbf{B}}[f_{i_T}]$ and $\widetilde{\boldsymbol{\Sigma}}[f_{i_T}]$ of the empirical pdf. The functionals are defined implicitly as in (4.176) as follows:

$$\mathbf{0} = \int_{\mathbb{R}^{N+K}} \boldsymbol{\psi}\left(\mathbf{x}, \mathbf{f}, \widetilde{\mathbf{B}}[h], \widetilde{\boldsymbol{\Sigma}}[h]\right) h(\mathbf{x}, \mathbf{f})\, d\mathbf{x} d\mathbf{f}. \tag{4.217}$$

The vector-valued function $\boldsymbol{\psi}$ in this expression follows from (4.214)-(4.215) and reads:

$$\boldsymbol{\psi}(\mathbf{x}, \mathbf{f}, \mathbf{B}, \boldsymbol{\Sigma}) \equiv \begin{pmatrix} w\left(\mathrm{Ma}^2(\mathbf{x}, \mathbf{B}\mathbf{f}, \boldsymbol{\Sigma})\right)\mathrm{vec}\left[(\mathbf{x} - \mathbf{B}\mathbf{f})\,\mathbf{f}'\right] \\ w\left(\mathrm{Ma}^2(\mathbf{x}, \mathbf{B}\mathbf{f}, \boldsymbol{\Sigma})\right)\mathrm{vec}\left[(\mathbf{x} - \mathbf{B}\mathbf{f})(\mathbf{x} - \mathbf{B}\mathbf{f})' - \boldsymbol{\Sigma}\right] \end{pmatrix}, \tag{4.218}$$

where vec is the operator $(A.104)$ that stacks the columns of a matrix into a vector. From (4.194) and (4.175) the norm of the influence function is proportional to the norm of the above vector:

$$\left\|\mathrm{IF}\left((\mathbf{x}, \mathbf{f}), f_{\mathbf{X}}, \left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)\right)\right\| \propto \|\boldsymbol{\psi}\|. \tag{4.219}$$

In particular, if the perturbations are normally distributed the term w in (4.218) becomes $w \equiv 1$, see (4.125). Therefore the influence function of the regression factor loadings estimator and the perturbation dispersion estimator is not bounded. This is not surprising, since we know from Section 4.3 that the ML estimators of the regression factor loadings of normally distributed factor models are the OLS coefficients, whose influence function is (4.212). In

other words, the ML estimator of the regression factor loadings in a factor model with normally distributed perturbations is not robust, and neither is the ML estimator of the perturbation dispersion.

On the other hand, if the perturbations are elliptically but not normally distributed the influence function display a different behavior. Consider for instance Cauchy-distributed perturbations. In this case as in (4.209) the term w in (4.218) becomes:

$$w(z) = \frac{N+1}{1+z}. \tag{4.220}$$

Therefore, from (4.219) and (4.202) the influence function becomes bounded. In other words, the ML estimators of the regression factor loadings and of the perturbation dispersion stemming from Cauchy-distributed perturbations are robust.

4.5.3 Robust estimators

From the above discussion we realize that robust estimators should satisfy two requirements. In the first place, since robustness questions the accuracy of the parametric assumptions on the unknown distribution of the invariants, the construction of robust estimators should be as independent as possible of these assumptions. Secondly, robust estimators should display a bounded influence function.

By forcing maximum likelihood estimators to have a bounded influence function, Maronna (1976) and Huber (1981) developed the so-called *M-estimators*, or generalized maximum likelihood estimators.

We recall that, under the assumption that the distribution of the market invariants is $f_{\boldsymbol{\theta}}$, the maximum likelihood estimators of the parameters $\boldsymbol{\theta}$ are defined as functional of the empirical distribution $\widetilde{\boldsymbol{\theta}}[f_{i_T}]$. From (4.176), this functional is defined as follows:

$$\widetilde{\boldsymbol{\theta}}[h]: \quad \int_{\mathbb{R}^N} \boldsymbol{\psi}\left(\mathbf{x}, \widetilde{\boldsymbol{\theta}}\right) h(\mathbf{x})\, d\mathbf{x} \equiv \mathbf{0}, \tag{4.221}$$

where $\boldsymbol{\psi}$ follows from the assumptions on the underlying distribution:

$$\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}) \equiv \frac{\partial \ln f_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \boldsymbol{\theta}}. \tag{4.222}$$

M-estimators are also defined by (4.221), but the function $\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})$ is chosen exogenously. Under these more general assumptions, the influence function (4.194) becomes:

$$\mathrm{IF}\left(\mathbf{x}, f_{\mathbf{X}}, \widehat{\boldsymbol{\theta}}\right) = \mathbf{A}\boldsymbol{\psi}\left(\mathbf{x}, \widetilde{\boldsymbol{\theta}}[f_{\mathbf{X}}]\right), \tag{4.223}$$

where the $S \times S$ matrix $\mathbf{A}$ is defined as follows:

$$\mathbf{A} \equiv -\left[\int_{\mathbb{R}^N} \left.\frac{\partial \boldsymbol{\psi}'}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}\equiv\widetilde{\boldsymbol{\theta}}[f_{\mathbf{X}}]} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\right]^{-1}, \tag{4.224}$$

see Appendix www.4.7.

This way the ensuing estimator $\widetilde{\boldsymbol{\theta}}[f_{i_T}]$ is independent of any assumption on the distribution of the underlying market invariants. If the function $\boldsymbol{\psi}$ is chosen appropriately, the influence function (4.223) becomes bounded. Therefore, the estimator $\widetilde{\boldsymbol{\theta}}[f_{i_T}]$ is robust.

Location and dispersion

Consider (4.207) and replace it with a vector-valued function $\boldsymbol{\psi}$ defined exogenously as follows:

$$\boldsymbol{\psi} \equiv \begin{pmatrix} \gamma\left(\mathrm{Ma}^2(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right)(\mathbf{x}-\boldsymbol{\mu}) \\ \zeta\left(\mathrm{Ma}^2(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})\right)\left(\mathrm{vec}\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'-\boldsymbol{\Sigma}\right)\right) \end{pmatrix}, \tag{4.225}$$

where γ and ζ are bounded functions that satisfy some regularity criteria. The ensuing estimators, which replace (4.203)-(4.205), solve the following implicit equations:

$$\widehat{\boldsymbol{\mu}} = \sum_{t=1}^{T} \frac{\gamma\left(\mathrm{Ma}^2\left(\mathbf{x}_t,\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)\right)}{\sum_{s=1}^{T}\gamma\left(\mathrm{Ma}^2\left(\mathbf{x}_s,\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)\right)}\mathbf{x}_t \tag{4.226}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t-\widehat{\boldsymbol{\mu}})(\mathbf{x}_t-\widehat{\boldsymbol{\mu}})'\,\zeta\left(\mathrm{Ma}^2\left(\mathbf{x}_t,\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}}\right)\right). \tag{4.227}$$

Since γ and ζ are bounded functions, so is the influence function and therefore these estimators are robust.

For instance, the following is a suitable choice of weights:

$$\gamma(x) \equiv \zeta(x) \equiv \begin{cases} 1 & \text{if } x \le x_0 \\ \frac{x_0}{x}e^{-\frac{(x-x_0)^2}{2b^2}} & \text{if } x > x_0, \end{cases} \tag{4.228}$$

where $x_0 \equiv \left(\sqrt{N}+2\right)/\sqrt{2}$. If we set $b \equiv +\infty$ we obtain the M-estimators suggested by Huber (1964). If we set $b \equiv 1.25$ we obtain the M-estimators suggested by Hampel (1973), see also Campbell (1980).

As in the case of the maximum likelihood estimators, in general the solution to the above implicit equations cannot be computed analytically. Nevertheless, for suitable choices of the functions γ and ζ such as (4.228) a recursive approach such as the following is guaranteed to converge. Further results for existence and uniqueness of the solution are provided in Huber (1981).

Step 0. Initialize $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ as the sample mean and sample covariance respectively.

Step 1. Compute the right hand side of (4.226) and (4.227).

Step 2. Update the left hand side of (4.226) and (4.227).

Step 3. If convergence has been reached stop, otherwise go to Step 1.

Explicit factors

It is possible to define multivariate M-estimators of the factor loadings and of the dispersion of the perturbations of an explicit factor model. The discussion proceeds exactly as above. Nevertheless, due to the larger number of parameters, convergence problems arise for the numerical routines that should yield the estimators in practice.

4.6 Practical tips

In this section we provide a few tips that turn out useful in practical estimation problems.

4.6.1 Detection of outliers

In Section 4.5.1 we introduced the tools to measure the effect on an estimate of one outlier both in the finite sample case, namely the influence curve and the jackknife, and in the infinite sample limit, namely the influence function. Another interesting question is the maximum amount of outliers that a certain estimator can sustain before breaking down: if there is a total of $T = T_G + T_O$ observations, where T_G are good data and T_O outliers, what is the highest ratio T_O/T that the estimator can sustain?

The *breakdown point* is the limit of this ratio when the number of observations tends to infinity. Obviously, the breakdown point is a positive number that cannot exceed 0.5.

For example, suppose that we are interested in estimating the location parameter of an invariant X_t.

Consider first the sample mean (4.41), which we report here:

$$\widehat{\mathrm{E}} \equiv \frac{1}{T}\sum_{t=1}^{T} x_t. \tag{4.229}$$

From (4.198) breakdown point of the sample mean is 0, as one single outlier can disrupt the estimation completely.

Consider now the sample median (4.39), which we report here:

$$\widehat{q}_{1/2} \equiv x_{[T/2]:T}, \tag{4.230}$$

where $[\cdot]$ denotes the integer part. The breakdown point of the median is 0.5. Indeed, changing the values of half the sample, i.e. all (minus one) the observations larger than $x_{[T/2]:T}$, or all (minus one) the observations smaller than $x_{[T/2]:T}$, does not affect the result of the estimation.

Estimators whose breakdown point is close to 0.5 are called *high breakdown estimators*. These estimators are useful in financial applications because they allow us to detect outliers. Indeed, time series are often fraught with suspicious data. In the case of one-dimensional variables it is relatively easy to spot these outliers by means of graphical inspection. In the multivariate case, this task becomes much more challenging.

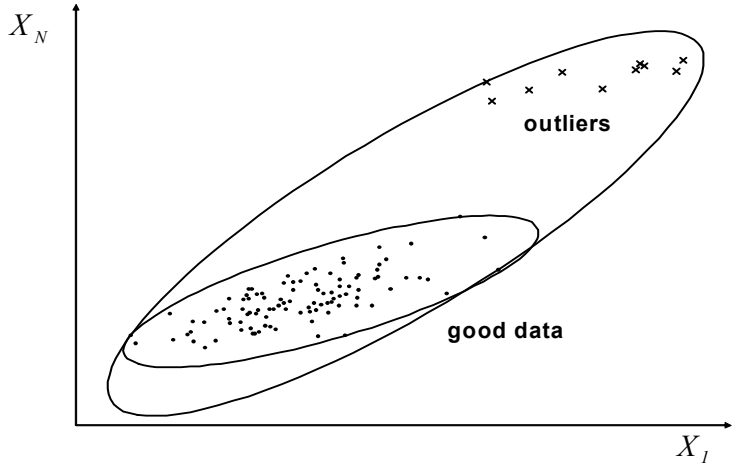

Fig. 4.19. Minimum Volume Ellipsoid

There exists a vast literature on estimators with high breakdown point, see Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986). Here we propose two methods to build high breakdown estimators of location and dispersion: the *minimum volume ellipsoid (MVE)* and the *minimum covariance determinant (MCD)*, see Rousseeuw and Leroy (1987), Rousseeuw and VanDriessen (1999). The rationale behind these estimators rests on the assumption that the core of the good data is tightly packed, whereas the joint set of good data and outliers is much more scattered, see Figure 4.19.

Minimum volume ellipsoid

Suppose we know that T_G out of the T data are good and T_O are outliers. Due to the above rationale, the smallest ellipsoid that contains the T_G good data is the smallest among all the ellipsoids that contain any set of T_G observations.

Consider a generic location parameter $\boldsymbol{\mu}$, i.e. an N-dimensional vector, and a generic scatter matrix $\boldsymbol{\Sigma}$, i.e. a positive and symmetric $N \times N$ matrix. The parameters $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ define an ellipsoid $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ as in (A.73). We can inflate this ellipsoid as follows:

$$\mathcal{E}^q_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \equiv \left\{\mathbf{x} \in \mathbb{R}^N \text{ such that } (\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \leq q^2\right\}. \tag{4.231}$$

This locus represents a rescaled version of the original ellipsoid, where all the principal axis are multiplied by a factor q. From $(A.77)$ the volume of the inflated ellipsoid reads:

$$\text{Vol}\left\{\mathcal{E}^q_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\right\} = \gamma_N q^N \sqrt{|\boldsymbol{\Sigma}|}, \tag{4.232}$$

where γ is the volume of the unit sphere:

$$\gamma_N \equiv \frac{\pi^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}+1\right)}. \tag{4.233}$$

Consider the set of Mahalanobis distances (2.61) of each observation from the location parameter $\boldsymbol{\mu}$ through the metric $\boldsymbol{\Sigma}$:

$$\text{Ma}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}_t \equiv \text{Ma}\left(\mathbf{x}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \equiv \sqrt{(\mathbf{x}_t-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t-\boldsymbol{\mu})}. \tag{4.234}$$

We can sort these distances in increasing order and consider the T_G-th distance:

$$q_{T_G} \equiv \text{Ma}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{T_G:T}. \tag{4.235}$$

By construction, the ellipsoid $\mathcal{E}^{q_{T_G}}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ contains only T_G points and from (4.232) its volume reads:

$$\text{Vol}\left\{\mathcal{E}^{q_{T_G}}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\right\} = \gamma_N \left(\text{Ma}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{T_G:T}\right)^N \sqrt{|\boldsymbol{\Sigma}|}. \tag{4.236}$$

Notice that the product on the right hand side of this expression does not depend on the determinant of $\boldsymbol{\Sigma}$. Therefore we can impose the constraint that the determinant of $\boldsymbol{\Sigma}$ be one.

Consequently, the parameters that give rise to the smallest ellipsoid that contains T_G observations solve the following equation:

$$\left(\widehat{\boldsymbol{\mu}}_{T_G}, \widehat{\boldsymbol{\Sigma}}_{T_G}\right) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}\succeq\mathbf{0},|\boldsymbol{\Sigma}|=1}{\text{argmin}} \left\{\text{Ma}^{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{T_G:T}\right\}, \tag{4.237}$$

where the the notation $\boldsymbol{\Sigma} \succeq \mathbf{0}$ means that $\boldsymbol{\Sigma}$ is symmetric and positive. Once we have computed the parameters (4.237), we tag as outliers all the observation that are not contained in the ellipsoid (4.231) determined by (4.237), with the radius (4.235) implied by (4.237).

In reality we do not know a priori the true number T_G of good data. Nevertheless, if T_G is the largest set of good data, the minimum volume ellipsoid that contains T_G+1 observations has a much larger volume than the minimum volume ellipsoid that contains T_G observations. Therefore, we consider the volume of the minimum volume ellipsoid as a function of the number of observations contained in the ellipsoid:

$$T_G \to \gamma_N \left(\mathrm{Ma}_{T_G:T}^{\widehat{\boldsymbol{\mu}}_{T_G}, \widehat{\boldsymbol{\Sigma}}_{T_G}} \right)^N . \tag{4.238}$$

The true number of good data is the value T_G where this function displays an abrupt jump, see Figure 4.20.

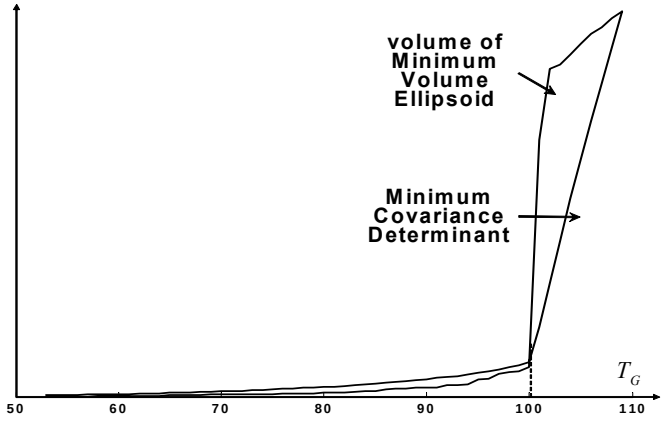

Fig. 4.20. Detection of outliers

The optimization problem (4.237) cannot be solved analytically. Numerical algorithms both deterministic and non-deterministic are available in the literature. We present below an approach that we used to generate the figures in this section.

Minimum covariance determinant

An alternative approach to detect outliers is provided by the minimum covariance determinant. This method also searches the "smallest ellipsoid". Instead of the smallest ellipsoid defined by the cloud of data we look for the smallest ellipsoid defined by the sample covariance of the data. Indeed, we recall from (4.48) that the sample covariance defines the smallest ellipsoid that fits the data in an average sense.

Suppose that we know the number of good observations T_G. Consider a generic subset of T_G observations $\{\mathbf{x}_1^*, \ldots, \mathbf{x}_{T_G}^*\}$ from the T observations in the time series i_T of the market invariants. We can compute the sample mean (4.41) and sample covariance (4.42) associated with this subset:

$$\widehat{\mathrm{E}}^*_{T_G} \equiv \frac{1}{T_G}\sum_{t=1}^{T_G}\mathbf{x}^*_t \tag{4.239}$$

$$\widehat{\mathrm{Cov}}^*_{T_G} \equiv \frac{1}{T_G}\sum_{t=1}^{T_G}\left(\mathbf{x}^*_t - \widehat{\mathrm{E}}^*_{T_G}\right)\left(\mathbf{x}^*_t - \widehat{\mathrm{E}}^*_{T_G}\right)'. \tag{4.240}$$

Consider the ellipsoid determined by these parameters:

$$\mathcal{E}^* \equiv \left\{\mathbf{x}:\ \left(\mathbf{x} - \widehat{\mathrm{E}}^*_{T_G}\right)'\left(\widehat{\mathrm{Cov}}^*_{T_G}\right)^{-1}\left(\mathbf{x} - \widehat{\mathrm{E}}^*_{T_G}\right) \leq 1\right\}. \tag{4.241}$$

From (A.77), the volume of $\mathcal{E}^*$ is proportional to the square root of the determinant of (4.240).

Therefore we have to determine the subset of observations that gives rise to the minimum covariance determinant:

$$\left\{\mathbf{x}^\times_1,\ldots,\mathbf{x}^\times_{T_G}\right\} = \underset{\mathbf{x}^*_1,\ldots,\mathbf{x}^*_{T_G}\in i_T}{\operatorname{argmin}}\left|\widehat{\mathrm{Cov}}^*_{T_G}\right|. \tag{4.242}$$

In reality we do not know a priori the true number T_G of good data. Nevertheless, if T_G is the largest set of good data, the minimum covariance determinant relative to T_G+1 observations is much larger than the minimum covariance determinant relative to T_G observations. Therefore, we consider the minimum covariance determinant as a function of the number of observations contained in the ellipsoid:

$$T_G \to \left|\widehat{\mathrm{Cov}}^\times_{T_G}\right|. \tag{4.243}$$

The true number of good data is the value T_G where this function displays an abrupt jump, see Figure 4.20.

The optimization problem (4.242) cannot be solved exactly. We present below an approach that we used to generate the figures in this section.

Computational issues

Suppose we have a series of observations $\{\mathbf{x}_1,\ldots,\mathbf{x}_T\}$. Assume we know that $T_G \leq T$ among them are good data.

In principle we should compute the minium volume ellipsoid and the sample covariance matrix for all the possible combinations of T_G observations out of the total T observations. This number reads:

$$\binom{T}{T_G} \equiv \frac{T_G!}{T_G!\,(T-T_G)!}, \tag{4.244}$$

which is intractably large if T exceeds the order of the dozen. Instead, we delete the unwelcome observations one at a time from the initial set of T observations using a theoretically sub-optimal, yet for practical purposes very effective, approach.

First we build Routine A, which computes the smallest ellipsoid $\mathcal{E}_{\mathbf{m},\mathbf{S}}$ that contains a given set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.

Step 0. Initialize the relative weights:

$$w_t \equiv \frac{1}{T}, \quad t = 1, \ldots, T. \tag{4.245}$$

Step 1. Compute the location parameter $\mathbf{m}$ and the scatter matrix $\mathbf{S}$ as follows:

$$\mathbf{m} \equiv \frac{1}{\sum_{s=1}^{T} w_s} \sum_{t=1}^{T} w_t \mathbf{x}_t \tag{4.246}$$

$$\mathbf{S} \equiv \sum_{t=1}^{T} w_t \left(\mathbf{x}_t - \mathbf{m}\right)\left(\mathbf{x}_t - \mathbf{m}\right)'. \tag{4.247}$$

Notice that the weights in the scatter matrix are *not* normalized.

Step 2. Compute the square Mahalanobis distances:

$$\mathrm{Ma}_t^2 \equiv \left(\mathbf{x} - \mathbf{m}\right)' \mathbf{S}^{-1} \left(\mathbf{x} - \mathbf{m}\right), \quad t = 1, \ldots, T. \tag{4.248}$$

Step 3. Update the weights: if $\mathrm{Ma}_t^2 > 1$ change the respective weight as follows:

$$w_t \mapsto w_t \, \mathrm{Ma}_t^2 ; \tag{4.249}$$

otherwise, leave the weight unchanged.

Step 4. If convergence has been reached, stop and define $\mathcal{E}_{\mathbf{m},\mathbf{S}}$ as in $(A.73)$, otherwise, go to Step 1.

Secondly, we build Routine B, which spots the farthest outlier in a series of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Define the following $T \times N$ matrix:

$$\mathbf{U} = \begin{pmatrix} \mathbf{x}_1' - \widehat{\mathrm{E}}' \\ \vdots \\ \mathbf{x}_T' - \widehat{\mathrm{E}}' \end{pmatrix}, \tag{4.250}$$

where $\widehat{\mathrm{E}}$ is the sample mean (4.41) of the data. The sample covariance matrix (4.42) can be written as follows:

$$\widehat{\mathrm{Cov}} \equiv \frac{1}{T} \mathbf{U}'\mathbf{U}. \tag{4.251}$$

We aim at finding the observation $\mathbf{x}_t$ such that if we remove it from the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ the determinant of the resulting sample covariance is reduced the most. This would mean that by dropping that observation the location-dispersion ellipsoid defined by sample mean and covariance shrinks the most, and thus that observation is the farthest outlier in the sample. To do this, we use the following result, see Poston, Wegman, Priebe, and Solka (1997):

$$\left|\mathbf{U}'_{(-t)}\mathbf{U}_{(-t)}\right| = (1-\lambda_t)\left|\mathbf{U}'\mathbf{U}\right|. \tag{4.252}$$

In this expression $\mathbf{U}_{(-t)}$ denotes the matrix (4.250) after removing the t-th row and λ_t denotes the t-th element of the diagonal of the *information matrix*:

$$\lambda_t \equiv \left(\mathbf{U}\left(\mathbf{U}'\mathbf{U}\right)^{-1}\mathbf{U}'\right)_{tt}. \tag{4.253}$$

It can be proved that

$$0 \leq \lambda_t \leq 1. \tag{4.254}$$

Therefore, the farthest outlier corresponds to the highest value of λ_t, unless $\lambda_t = 1$: in this last case, if we remove the t-th observation the sample covariance becomes singular, as is evident from (4.252).

Now we can define Routine C, which detects the outliers among the given data by means of the minimum volume ellipsoid and the minimum covariance determinant.

Step 0. Consider as data all the observations.

Step 1. Compute the sample mean and covariance $\left(\widehat{\mathrm{E}}, \widehat{\mathrm{Cov}}\right)$ of the given data and compute the determinant of the sample covariance $\left|\widehat{\mathrm{Cov}}\right|$.

Step 2. Compute with Routine A the minimum volume ellipsoid of the given data $\mathcal{E}_{\mathbf{m},\mathbf{S}}$ and compute $|\mathbf{S}|$.

Step 3. Find the farthest outlier among the data with Routine B and remove it from the data.

Step 4. If the number of data left is less than half the original number stop, otherwise go to Step 1.

The plot of $\left|\widehat{\mathrm{Cov}}\right|$ and/or $|\mathbf{S}|$ as a function of the number of observations in the dataset shows an abrupt jump when the first outlier is added to the dataset, see Figure 4.20. The respective sample covariance $\widehat{\mathrm{Cov}}$ is the minimum covariance determinant and the respective ellipsoid $\mathcal{E}_{\mathbf{m},\mathbf{S}}$ is the minimum volume ellipsoid.

4.6.2 Missing data

Sometimes some data is missing from the time series of observations. Our purpose is twofold. On the one hand, we are interested in interpolating the missing values. On the other hand we want to estimate parameters of interest regarding the market invariants, such as parameters of location or dispersion. We refer the reader to Stambaugh (1997) for a discussion of the case where some series are shorter than others. Here, we discuss the case where some observations are missing randomly from the time series.

Consider a $T \times N$ panel of observations, where T is the length of the sample and N is the number of market invariants. Each row of this matrix corresponds to a joint observation $\mathbf{x}_t$ of the invariants at a specific date. In some rows one or more entry might be missing:

$$\mathbf{x}_t \equiv \mathbf{x}_{t,\text{mis}(t)} \cup \mathbf{x}_{t,\text{obs}(t)}, \tag{4.255}$$

where we stressed that the set of missing and observed values depends on the specific date t. Notice that for most t the set $\mathbf{x}_{t,\text{mis}(t)}$ is empty.

For example, consider a case of four market invariants and a hundred joint observations. Assume that the second entry is missing at time $t = 7$ then

$$\text{mis}\,(7) \equiv \{2\}\,, \quad \text{obs}\,(7) \equiv \{1,3,4\}\,. \tag{4.256}$$

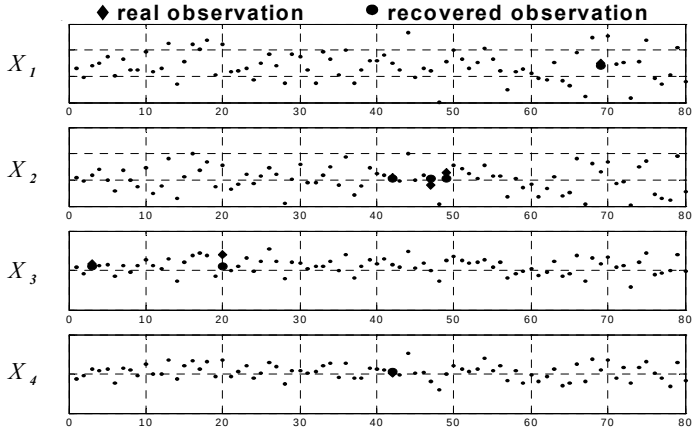

Fig. 4.21. EM algorithm for data recovery

Following Little and Rubin (1987) we make the simplifying assumption that prior to their realization the invariants are independent and normally distributed:

$$\begin{pmatrix} \mathbf{X}_{t,\text{mis}(t)} \\ \mathbf{X}_{t,\text{obs}(t)} \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{\text{mis}(t)} \\ \boldsymbol{\mu}_{\text{obs}(t)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\text{mis}(t),\text{mis}(t)} & \boldsymbol{\Sigma}_{\text{mis}(t),\text{obs}(t)} \\ \boldsymbol{\Sigma}_{\text{obs}(t),\text{mis}(t)} & \boldsymbol{\Sigma}_{\text{obs}(t),\text{obs}(t)} \end{pmatrix} \right). \tag{4.257}$$

The algorithm we propose is a specific instance of a general approach called *expectation-maximization (EM) algorithm*, see Dempster, Laird, and Rubin (1977) and also Bilmes (1998). In Figure 4.21 we recovered a few missing values with the EM algorithm.

The algorithm proceeds as follows, see Appendix www.4.8 for the proofs.

Step 0. Set $u \equiv 0$ and initialize both the location and the dispersion parameters. For all $n = 1 \ldots, N$ set:

$$\mu_n^{(u)} \equiv \frac{1}{T_n} \sum_{t \in \text{ avail. obs.}} x_{t,n} \tag{4.258}$$

$$\Sigma_{nn}^{(u)} \equiv \frac{1}{T_n} \sum_{t \in \text{ avail. obs.}} \left(x_{t,n} - \mu_n^{(u)}\right)^2, \tag{4.259}$$

where T_n is the number of available observations for the generic n-th market invariant. For all $n, m = 1 \ldots, N$, $n \neq m$ set:

$$\Sigma_{nm}^{(u)} \equiv 0. \tag{4.260}$$

Step 1. For each $t = 1, \ldots, T$ fill in the missing entries by replacing the missing values with their expected value conditional on the observations. For the observed values we have:

$$\mathbf{x}_{t,\text{obs}(t)}^{(u)} \equiv \mathbf{x}_{t,\text{obs}(t)}; \tag{4.261}$$

and for the missing values we have:

$$\begin{aligned} \mathbf{x}_{t,\text{mis}(t)}^{(u)} \equiv & \boldsymbol{\mu}_{\text{mis}(t)}^{(u)} \\ & + \boldsymbol{\Sigma}_{\text{mis}(t),\text{obs}(t)}^{(u)} \left(\boldsymbol{\Sigma}_{\text{obs}(t),\text{obs}(t)}^{(u)}\right)^{-1} \left(\mathbf{x}_{t,\text{obs}(t)} - \boldsymbol{\mu}_{\text{obs}(t)}^{(u)}\right). \end{aligned} \tag{4.262}$$

Step 2. For each $t = 1, \ldots, T$ compute the conditional covariance, which is zero if at least one of the invariants is observed:

$$\mathbf{C}_{t,\text{obs}(t),\text{mis}(t)}^{(u)} \equiv \mathbf{0}, \quad \mathbf{C}_{t,\text{obs}(t),\text{obs}(t)}^{(u)} \equiv \mathbf{0}, \tag{4.263}$$

and otherwise reads:

$$\begin{aligned} \mathbf{C}_{t,\text{mis}(t),\text{mis}(t)}^{(u)} \equiv & \boldsymbol{\Sigma}_{\text{mis}(t),\text{mis}(t)}^{(u)} \\ & - \boldsymbol{\Sigma}_{\text{mis}(t),\text{obs}(t)}^{(u)} \left(\boldsymbol{\Sigma}_{\text{obs}(t),\text{obs}(t)}^{(u)}\right)^{-1} \boldsymbol{\Sigma}_{\text{obs}(t),\text{mis}(t)}^{(u)}. \end{aligned} \tag{4.264}$$

Step 3. Update the estimate of the location parameter:

$$\boldsymbol{\mu}^{(u+1)} \equiv \frac{1}{T} \sum_t \mathbf{x}_t^{(u)}. \tag{4.265}$$

Step 4. Update the estimate of the dispersion parameter:

$$\boldsymbol{\Sigma}^{(u+1)} \equiv \frac{1}{T} \sum_t \mathbf{C}_t^{(u)} - \left(\boldsymbol{\mu}^{(u+1)} - \boldsymbol{\mu}^{(u)}\right)\left(\boldsymbol{\mu}^{(u+1)} - \boldsymbol{\mu}^{(u)}\right)'. \tag{4.266}$$

Step 5. If convergence has been reached, stop. Otherwise, set $u \equiv u + 1$ and go to Step 1.

4.6.3 Weighted estimates

We have seen in (4.167) and comments that follow that any estimator $\widehat{\mathbf{G}}$ can be represented as a functional $\widetilde{\mathbf{G}}[f_{i_T}]$ that acts on the empirical probability density function of the time series of the market invariants:

$$i_T \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}. \tag{4.267}$$

In the definition of the empirical density function (4.168) and thus in the definition of the estimator $\widehat{\mathbf{G}}$ the order of the realization of the market invariants does not play a role. This is correct, since the invariants are independent and identically distributed across time, see (4.5).

Nevertheless, intuition suggests that the most recent observations should somehow play a more important role than observations farther back in the past. To account for this remark, it suffices to replace the definition of the empirical probability density function (4.168) as follows:

$$f_{i_T} \mapsto f_{i_T} \equiv \frac{1}{\sum_{s=1}^{T} w_s} \sum_{t=1}^{T} w_t \delta^{(\mathbf{x}_t)}, \tag{4.268}$$

where δ is the Dirac delta ($B.17$) and where the weights w_t are positive, non-decreasing functions of the time index t. We present below two notable cases.

Rolling window

A simple way to give more weight to the last observations is to assume that only the last set of observations is good at forecasting, whereas considering the previous observations might be disruptive. Therefore, we consider only the *rolling window* of the last W observations among the T in the whole time series. This corresponds to setting in (4.268) the following weights:

$$w_t \equiv 1, \text{ if } t > T - W \tag{4.269}$$

$$w_t \equiv 0, \text{ if } t \leq T - W. \tag{4.270}$$

Each time a new observation is added to the time series, we roll over the window and again we only consider the last W observations.

For example, if we are at time T, the sample mean (4.41) becomes:

$$\widehat{\mathrm{E}}_W \equiv \frac{1}{W} \sum_{t=T-W+1}^{T} \mathbf{x}_t, \tag{4.271}$$

and the sample covariance (4.42) becomes:

$$\widehat{\mathrm{Cov}}_W \equiv \frac{1}{W} \sum_{t=T-W+1}^{T} (\mathbf{x}_t - \widehat{\boldsymbol{\mu}}_w)(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}_w)'. \tag{4.272}$$

To determine the most appropriate value of the rolling window one should keep in mind the specific investment horizon.

Exponential smoothing

A less dramatic approach consists in giving less and less weight to past observations in a smooth fashion. The *exponential smoothing* consists in setting in (4.268) weights that decay exponentially:

$$w_t \equiv (1-\lambda)^{T-t}, \tag{4.273}$$

where λ is a fixed *decay factor* between zero and one. Notice that the case $\lambda \equiv 0$ recovers the standard empirical pdf. If the decay factor is strictly positive, the weight of past observations in the estimate tapers at an exponential rate.

For example, if we are at time T, the sample mean (4.41) becomes:

$$\widehat{\mathrm{E}}_\lambda \equiv \frac{\lambda}{1-(1-\lambda)^T} \sum_{t=1}^{T} (1-\lambda)^{T-t} \mathbf{x}_t; \tag{4.274}$$

and the sample covariance (4.42) becomes:

$$\widehat{\mathrm{Cov}}_\lambda \equiv \frac{\lambda}{1-(1-\lambda)^T} \sum_{t=1}^{T} (1-\lambda)^{T-t} \left(\mathbf{x}_t - \widehat{\mathrm{E}}_\lambda\right)\left(\mathbf{x}_t - \widehat{\mathrm{E}}_\lambda\right)'. \tag{4.275}$$

The exponential smoothing estimate is used, among others, by RiskMetrics and Goldman Sachs, see Litterman and Winkelmann (1998). To assign a suitable value to the decay factor a possible approach is to choose a parametric form for the probability density function and then apply the maximum likelihood principle (4.66).

For example, if the invariants $\mathbf{X}_t$ are normally distributed, we determine the parameter λ in (4.274)-(4.275) by maximizing the normal log-likelihood:

$$\widetilde{\lambda} \equiv \underset{0\leq\lambda<1}{\operatorname{argmax}} \left(-\frac{T}{2} \ln \left|\widehat{\mathrm{Cov}}_\lambda\right| - \frac{1}{2} \sum_{t=1}^{T} \left(\mathbf{x}_t - \widehat{\mathrm{E}}_\lambda\right)' \widehat{\mathrm{Cov}}_\lambda^{-1} \left(\mathbf{x}_t - \widehat{\mathrm{E}}_\lambda\right) \right). \tag{4.276}$$

The exponential smoothing presents an interesting link to *GARCH* models, an acronym for Generalized AutoRegressive Conditionally Heteroskedastic models, see Engle (1982) and Bollerslev (1986). Indeed, by recursive substitution we can check that in the presence of an infinite series of observations the exponential smoothing is consistent with the following GARCH model:

$$X_t \equiv \mu + \epsilon_t, \tag{4.277}$$

where ϵ_t are random perturbations such that:

$$\mathrm{Var}\{\epsilon_t\} = \lambda \epsilon_{t-1}^2 + (1-\lambda)\,\mathrm{Var}\{\epsilon_{t-1}\}. \tag{4.278}$$

4.6.4 Overlapping data

In order for the market invariants to be independent across time it is necessary that they refer to non-overlapping time intervals as in Figure 3.11.

For example, consider the case of the equity market where the invariants are the compounded returns (3.11). Suppose that the returns are identically normally distributed and independent:

$$\begin{pmatrix} C_{t,\tau} \\ C_{t+\tau,\tau} \\ \vdots \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu \\ \mu \\ \vdots \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \ddots \\ 0 & \sigma^2 & \ddots \\ \ddots & \ddots & \ddots \end{pmatrix} \right). \tag{4.279}$$

From (2.163) we immediately derive the distribution of the overlapping time series:

$$\begin{pmatrix} C_{t,2\tau} \\ C_{t+\tau,2\tau} \\ \vdots \end{pmatrix} = \begin{pmatrix} C_{t,\tau} + C_{t-\tau,\tau} \\ C_{t+\tau,\tau} + C_{t,\tau} \\ \vdots \end{pmatrix} \sim \mathrm{N}(\mathbf{m}, \mathbf{S}), \tag{4.280}$$

where

$$\mathbf{m} \equiv \begin{pmatrix} 2\mu \\ 2\mu \\ \vdots \end{pmatrix}, \quad \mathbf{S} \equiv \begin{pmatrix} 2\sigma^2 & \sigma^2 & \cdots \\ \sigma^2 & 2\sigma^2 & \ddots \\ \ddots & \ddots & \ddots \end{pmatrix}. \tag{4.281}$$

This expression shows that that the overlapping observations are not independent.

In some circumstances it is possible and even advisable to consider overlapping data, see Campbell, Lo, and MacKinlay (1997).

4.6.5 Zero-mean invariants

When a location parameter such as the expected value of a market invariant is close to null with respect to a dispersion parameter such as its standard deviation, it might be convenient to assume that the location parameter is zero, instead of estimating it. We can interpret this approach as an extreme case of shrinkage, see Section 4.4.

This approach often leads to better results, see Alexander (1998) and therefore it is often embraced by practitioners. For instance we made this assumption in (3.233) regarding the expected changes in yield in the swap market.

4.6.6 Model-implied estimation

Time-series analysis is by definition backward-looking. An alternative approach to estimation makes use of pricing models, which reflects the expectations on the market and thus is forward-looking.

Consider a parametric model $f_{\boldsymbol{\theta}}$ for the market invariants. Assume there exist pricing functions $\mathbf{F}(\boldsymbol{\theta})$ of financial products which depend on those parameters and which trade at the price $\mathbf{P}_T$ at the time the estimate is made.

In these circumstances we can compute the estimate of the parameters as the best fit to the data, i.e. as the solution of the following optimization problem:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ (\mathbf{P} - \mathbf{F}(\boldsymbol{\theta}))' \mathbf{Q} (\mathbf{P} - \mathbf{F}(\boldsymbol{\theta})) \right\}, \tag{4.282}$$

where $\mathbf{Q}$ is a suitably chosen symmetric and positive matrix.

Depending on the applications, some authors suggest mixed approaches, where time series analysis is used together with implied estimation.

For example, to estimate the correlation matrix of swap yield changes we can proceed as in Longstaff, Santa-Clara, and Schwartz (2001). First we estimate from (4.41) and (4.42) the sample correlation matrix:

$$\widehat{C}_{mn} \equiv \frac{\widehat{\operatorname{Cov}}\{X_m, X_n\}}{\sqrt{\widehat{\operatorname{Cov}}\{X_m, X_m\}\,\widehat{\operatorname{Cov}}\{X_n, X_n\}}}. \tag{4.283}$$

Then we perform the principal component decomposition $(A.70)$ of the correlation matrix:

$$\widehat{\mathbf{C}} = \widehat{\mathbf{E}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{E}}', \tag{4.284}$$

where $\widehat{\boldsymbol{\Lambda}}$ is the diagonal matrix of the estimated eigenvalues and $\widehat{\mathbf{E}}$ is the orthogonal matrix of the respective estimated eigenvectors. Next, we assume that a more suitable estimate of the correlation matrix is of this form:

$$\mathbf{C} = \widehat{\mathbf{E}}\boldsymbol{\Psi}\widehat{\mathbf{E}}', \tag{4.285}$$

where $\boldsymbol{\Psi}$ is a diagonal matrix of positive entries. Finally we fit an estimate $\widetilde{\boldsymbol{\Psi}}$ from the prices of a set of swaptions which depend on the correlation through suitable pricing function.

The main problem with the model-implied approach is that the pricing functions $\mathbf{F}(\boldsymbol{\theta})$ give rise to model risk. This risk is equivalent to the risk

of assuming an incorrect parametric distribution for the invariants in the derivation of maximum likelihood estimators.