6. Black-Litterman Optimization

Prof. Gordon Ritter

October 15, 2017

6.1. **Statistics Review.** At a high level, forecasting may be viewed as a three-step procedure:

(1) Build a model for the quantity you want to forecast, which connects that quantity to some parameters $\boldsymbol{\theta}$.
(2) Learn all you can about the parameter $\boldsymbol{\theta}$ from the data you have (perhaps using additional models which connect $\boldsymbol{\theta}$ to other quantities you can observe)
(3) Use the model(s), and what you have learned about the parameter, to calculate the expected value of the quantity you want to forecast.

Here "what you have learned about the parameter" is encoded as a distribution on parameter space. This distribution is called the *posterior*.

For any probability measure p and any events A, B,

$$\begin{aligned} p(A \,|\, B)p(B) &= p(A \cap B) = p(B \cap A) = p(B \,|\, A)p(A) \\ p(A \,|\, B) &= p(B \,|\, A)\frac{p(A)}{p(B)}\,. \end{aligned} \tag{6.1}$$

This result, known as *Bayes' theorem*, appears trivial but historically was a major conceptual step because it allows the "inversion" of probabilities.

Eq. (6.1) can be applied to any events A and B, but is most usefully applied when A is the event of observing a certain data set (usually the one given to us by the world), and B is the event of parameters taking on certain values.

In language universally familiar to statisticians (Robert, 2007), a *Bayesian statistical model* consists of:

(1) a vector-valued random variable $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ distributed according to $f(\mathbf{x} \,|\, \boldsymbol{\theta})$, where realizations of $\mathbf{x}$ have been observed and only the parameter $\boldsymbol{\theta}$ (which belongs to a real vector space $\Theta \subseteq \mathbb{R}^\ell$) is unknown, and
(2) a prior density $\pi(\boldsymbol{\theta})$ on Θ.

The function $f(\mathbf{x} \,|\, \boldsymbol{\theta})$ is called the *likelihood* and, after conditioning on $\boldsymbol{\theta}$, forms a density on the *data space* $\mathcal{X} \subseteq \mathbb{R}^d$. Applying Bayes' theorem to the likelihood function gives the *posterior*

$$p(\boldsymbol{\theta} \,|\, \mathbf{x}) = \frac{f(\mathbf{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization}}$$

The posterior is best viewed as the density on Θ proportional to $f(\mathbf{x} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$, and the normalization factor $p(\mathbf{x})$ drops out of certain calculations. In Bayesian statistics, all statistical inference is based on the posterior.

In particular, given any quantity of interest $\boldsymbol{r}$ which is related to $\boldsymbol{\theta}$ by some other density $p(\boldsymbol{r} \,|\, \boldsymbol{\theta})$, with the posterior in hand we can form a prediction of $\boldsymbol{r}$:

$$\mathbb{E}[\boldsymbol{r} \,|\, \mathbf{x}] = \int \boldsymbol{r} \underbrace{\int p(\boldsymbol{r} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{x}) d\boldsymbol{\theta}}_{\text{posterior predictive density}} d\boldsymbol{r}$$

6.2. **Black, Litterman, and Bayes.** The topic of portfolio optimization in the style of Black and Litterman (1991) and Black and Litterman (1992) seems to have generated more than its share of confusion over the years, as evidenced by articles with titles such as *"A demystification of the Black–Litterman model"* (Satchell and Scowcroft, 2000), etc. The method itself is often described as "Bayesian" but the original authors do not elaborate directly on connections with Bayesian statistics. The paper by He and Litterman (1999) contains many references to a "prior" but only one mention of a "posterior" without details, and no mention of a "likelihood."

In the present lecture, we clarify the exact nature of the Bayesian statistical model to which Black-Litterman optimization corresponds, in terms of the prior, likelihood, and posterior. In the process we also lay out the full set of assumptions made, some of which are glossed over in other treatments.

Consider a view such as "the German equity market will outperform a capitalization-weighted basket of the rest of the European equity markets by 5%," which is an example presented in He and Litterman (1999). Let $\boldsymbol{p} \in \mathbb{R}^n$ denote a portfolio which is long one unit of the DAX index, and short a one-unit basket which holds each of the other major European indices (UKX, CAC40, AEX, etc.) in proportion to their respective aggregate market capitalizations, so that $\sum_i p_i = 0$. Let $q = 0.05$ in this example. This view may be equivalently expressed as

$$\mathbb{E}[\boldsymbol{p}'\boldsymbol{r}] = q \in \mathbb{R} \tag{6.2}$$

where $\boldsymbol{r}$ is the random vector of asset returns over some subsequent interval, and q denotes the expected return, according to the view.

If there are multiple such views, say

$$\mathbb{E}[\boldsymbol{p}_i'\boldsymbol{r}] = q_i, \quad i = 1 \dots k$$

then the portfolios $\boldsymbol{p}_i$ are more conveniently arranged as rows of a matrix $\boldsymbol{P}$, and the statement of views becomes

$$\mathbb{E}[\boldsymbol{P}\boldsymbol{r}] = \boldsymbol{q} \text{ for } \boldsymbol{q} \in \mathbb{R}^k. \tag{6.3}$$

In the language of statistics, the core idea of Black and Litterman (1991) is to treat the portfolio manager's views as noisy observations which are useful for performing statistical inference concerning the parameters in some underlying model for $\boldsymbol{r}$. For example, if

$$\boldsymbol{r} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{6.4}$$

with $\boldsymbol{\Sigma}$ a known positive-definite $n \times n$ matrix, then the views (6.3) can be recast as "observations" relevant for inference on the parameter $\boldsymbol{\theta}$.

The practitioner must also specify a level of uncertainty or "error bar" for each view, which is assumed to be an independent source of noise from the volatility already accounted for in a model such as (6.4). This is expressed as the following more precise restatement of (6.3):

$$\boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{q} + \boldsymbol{\epsilon}^{(v)}, \quad \boldsymbol{\epsilon}^{(v)} \sim N(0, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \dots, \omega_k) \tag{6.5}$$

Eq. (6.5) specifies a Gaussian distribution and may be alternately written in the form of a likelihood:

$$f(\boldsymbol{q} \,|\, \boldsymbol{\theta}) \propto \exp\big[-\frac{1}{2}(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q})'\boldsymbol{\Omega}^{-1}(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q})\big] \tag{6.6}$$

which is the standard normal likelihood for a multiple linear regression problem with dependent variable $\boldsymbol{q}$ and design matrix $\boldsymbol{P}$.

Portfolio managers in this model specify *noisy, partial, indirect* information about $\boldsymbol{\theta}$, through their views. The information is *partial and indirect* because the views are on portfolio returns, i.e. linear transformations of returns, rather than on the asset returns directly. The information is *noisy*, with the noise modeled by $\boldsymbol{\epsilon}^{(v)}$, because the future is always uncertain.

A subjective, uncertain view about what will happen to a certain portfolio in the future is conceptually distinct from a noisy experimental observation such as an attempt to measure some physical constant with imperfect laboratory equipment. Nonetheless, for building intuition, we suggest thinking of a portfolio manager's forecast as an "observation of the future" in which the measuring device is a rather murky and unreliable crystal ball. Only in this way is it analogous to the noisy measurements in experimental design which much of statistics is designed to model.

Quite generally, if any random variable r comes from a density $p(r \,|\, \theta)$ with parameter θ, and if one were given a set of noisy observations of realizations of r, then one could infer something about θ by statistical inference. This would be the predicament of a physicist with a noisy measuring device, measuring a quantity that is itself random, and we suppose the physicist wants to know about the underlying data-generating process. Black and Litterman essentially say that the portfolio manger's view is, mathematically, no different from a noisy observation of a realization of (a linear transformation of) future returns.

As noted above, to perform statistical inference, observations alone are not sufficient; one needs to fully specify the statistical model, which includes a likelihood and a prior. In fact (6.5) specifies the likelihood as

$$f(\boldsymbol{q} \,|\, \boldsymbol{\theta}) \propto \exp\big[-\frac{1}{2}(\boldsymbol{P\theta} - \boldsymbol{q})'\boldsymbol{\Omega}^{-1}(\boldsymbol{P\theta} - \boldsymbol{q})\big] \tag{6.7}$$

which is the standard normal likelihood for a multiple linear regression problem with dependent variable $\boldsymbol{q}$ and design matrix $\boldsymbol{P}$.

A feature of Bayesian statistics that is dissimilar from frequentist statistics is the ability to perform inference in data-scarce situations. In Bayesian statistics, even a single observation can lead to valid inferences for multiparameter models due to the presence of a prior. In essence, when less information is available, more weight is given to the prior.

The classic regression problem has the number of variables much less than the number of observations, and is therefore identifiable. However, the need to perform inference in models with many more variables than observations also arises in many applications. Notably, this arises in the analysis of gene expression arrays, and is typically handled by Bayesian methods such as ridge and the lasso (Tibshirani, 1996).

In a Black-Litterman model with one single view, there is one observation and still n parameters to serve as the subjects for statistical inference: $\boldsymbol{\theta} \in \mathbb{R}^n$ are the unobservable means of the asset returns. More generally, we may be presented with no views, one, or very many. When views are collected from many diverse portfolio managers or economists, they may contain internal contradictions; ie. it may be impossible that they all come true exactly. Bayesian regression is the ideal tool to deal with all such cases. Internal contradictions in the views simply mean that there is no exact (zero-residual) solution to the regression equations, which in fact is the typical situation in classic (identifiable) linear regression.

We have not yet specified the prior, but Black and Litterman were motivated by the guiding principle that, in the absence of any sort of information/views which could constitute alpha over the benchmark, the optimization procedure should simply return the global CAPM equilibrium portfolio, with holdings denoted $\boldsymbol{h}_{eq}$. Hence in the absence of any views, and with prior mean equal to $\boldsymbol{\Pi}$, the investor's model of the world is that

$$\boldsymbol{r} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \quad \text{and} \quad \boldsymbol{\theta} \sim N(\boldsymbol{\Pi}, \boldsymbol{C}) \tag{6.8}$$

for some covariance $\boldsymbol{C}$ whose inverse represents the amount of precision in the prior. For any portfolio $\boldsymbol{p}$, then, according to (6.8) we have

$$\mathbb{E}[\boldsymbol{p}'\boldsymbol{r}] = \boldsymbol{p}'\boldsymbol{\Pi} \quad \mathbb{V}[\boldsymbol{p}'\boldsymbol{r}] = \boldsymbol{p}'(\boldsymbol{\Sigma} + \boldsymbol{C})\boldsymbol{p}\,. \tag{6.9}$$

In fact we must make a choice whether to use the conditional or unconditional variance in optimization: $\mathbb{V}(\boldsymbol{r} \,|\, \boldsymbol{\theta}) = \boldsymbol{\Sigma}$ but $\mathbb{V}(\boldsymbol{r}) = \boldsymbol{\Sigma} + \boldsymbol{C}$. Since investors are presumably concerned with unconditional variance of wealth, the unconditional variance form is preferable, but when we take $\boldsymbol{C} = \tau\boldsymbol{\Sigma}$, this distinction won't matter much (it amounts to a different risk-aversion constant).

Mean-variance optimization with (6.9) and with risk-aversion parameter $\delta > 0$, leads to $\delta^{-1}(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}\boldsymbol{\Pi}$ as the optimal portfolio; setting this equal to the CAPM equilibrium then gives

$$\boldsymbol{h}_{\text{eq}} = \delta^{-1}(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}\boldsymbol{\Pi}. \tag{6.10}$$

Any combination of $\boldsymbol{\Pi}, \boldsymbol{C}$ satisfying (6.10) will lead to a model with the desired property – that the optimal portfolio with only the information given in the prior is the prescribed portfolio $\boldsymbol{h}_{\text{eq}}$. In particular, taking $\boldsymbol{C} = \tau\boldsymbol{\Sigma}$ with some arbitrary scalar $\tau > 0$, as did the original authors, leads to

$$\boldsymbol{\Pi} = \delta(1 + \tau)\boldsymbol{\Sigma}\boldsymbol{h}_{\text{eq}}$$

We thus have the normal likelihood (6.7) and the normal prior (6.8) which is a *conjugate prior* for that likelihood, meaning that the posterior is of the same family (ie. also normal in this example). A detailed discussion of conjugate priors is found in Robert (2007, Sec. 3.3).

The negative log posterior is thus proportional to (neglecting terms that don't contain $\boldsymbol{\theta}$):

$$(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q})'\boldsymbol{\Omega}^{-1}(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q}) + (\boldsymbol{\theta} - \boldsymbol{\Pi})'\boldsymbol{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\Pi}) \tag{6.11}$$

$$= \boldsymbol{\theta}'\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{\theta}'\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q} - \boldsymbol{q}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}\boldsymbol{\theta} \tag{6.12}$$

$$+ \boldsymbol{\theta}'\boldsymbol{C}^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}'\boldsymbol{C}^{-1}\boldsymbol{\Pi} - \boldsymbol{\Pi}'\boldsymbol{C}^{-1}\boldsymbol{\theta}$$

$$= \boldsymbol{\theta}'[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}]\boldsymbol{\theta} - 2(\boldsymbol{q}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{\Pi}'\boldsymbol{C}^{-1})\boldsymbol{\theta} \tag{6.13}$$

Recall that, for $\boldsymbol{H}$ symmetric,

$$\boldsymbol{\theta}'\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{v}'\boldsymbol{H}\boldsymbol{\theta} = (\boldsymbol{\theta} - \boldsymbol{v})'\boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{v}) - \boldsymbol{v}'\boldsymbol{H}\boldsymbol{v}$$

Therefore if a multivariate normal random variable $\boldsymbol{\theta}$ has density $p(\boldsymbol{\theta})$,

$$\begin{aligned} -2\log p(\boldsymbol{\theta}) \quad &= \quad \boldsymbol{\theta}'\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\eta}'\boldsymbol{\theta} + (\text{terms without } \boldsymbol{\theta}) \\ &\Rightarrow \ \mathbb{V}[\boldsymbol{\theta}] = \boldsymbol{H}^{-1} \text{ and } \mathbb{E}\boldsymbol{\theta} = \boldsymbol{H}^{-1}\boldsymbol{\eta}. \end{aligned}$$

For the quadratic term to match (6.13) we must have $\boldsymbol{H} = \boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}$ and hence the posterior has mean

$$\boldsymbol{v} = [\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}]^{-1}[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q} + \boldsymbol{C}^{-1}\boldsymbol{\Pi}] \tag{6.14}$$

and covariance

$$\boldsymbol{H}^{-1} = [\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}]^{-1}. \tag{6.15}$$

Part of the beauty of this derivation is its simplicity: going from (6.11) to (6.15) requires just a few lines of algebra.

Investors with CARA utility of final wealth will want to solve

$$\boldsymbol{h}^* = \underset{\boldsymbol{h}}{\operatorname{argmax}} \left\{ \mathbb{E}[\boldsymbol{h}'\boldsymbol{r}] - (\delta/2)\mathbb{V}[\boldsymbol{h}'\boldsymbol{r}] \right\}$$

where $\mathbb{E}[\boldsymbol{r}]$ and $\mathbb{V}[\boldsymbol{r}]$ denote, respectively, the unconditional mean and covariance of $\boldsymbol{r}$ under the posterior. The unconditional covariance is a sum of variance due to parameter uncertainty, and variance due to the randomness in $\boldsymbol{r}$. In other words,

$$\mathbb{V}[\boldsymbol{h}'\boldsymbol{r}] = \boldsymbol{h}'[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}]^{-1}\boldsymbol{h} + \boldsymbol{h}'\boldsymbol{\Sigma}\boldsymbol{h}$$

The optimal portfolio accounting for both types of variance is then

$$\boldsymbol{h}^* = \delta^{-1}[\boldsymbol{H}^{-1} + \boldsymbol{\Sigma}]^{-1}\boldsymbol{H}^{-1}[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q} + \boldsymbol{C}^{-1}\boldsymbol{\Pi}]\,.$$

6.3. **Generalizing the Model.** The observations in the previous section now allow us to easily formulate the most general model of this type.

Definition 6.1. A *Black-Litterman-Bayes* model consists of:

(a) A parametric statistical model for asset returns $p(\boldsymbol{r} \,|\, \boldsymbol{\theta})$ with finite-dimensional parameter vector $\boldsymbol{\theta}$,
(b) A prior $\pi(\boldsymbol{\theta})$ on the parameter space,
(c) A likelihood function $f(\boldsymbol{q} \,|\, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is any parameter vector appearing in a parametric statistical model for asset returns, and $\boldsymbol{q}$ is a vector supplied by portfolio managers or economists.
(d) A utility function $u(w)$ of final wealth in the sense of Arrow (1971) and Pratt (1964).

Items (a)-(b) simply state that we have a Bayesian statistical model, as defined above, for asset returns. Under such a model, Decision Theory (see Robert (2007, Ch. 2) and references) teaches us that the optimal decision is the one maximizing posterior expected utility. This leads us to Definition 6.2.

Definition 6.2. Given a Black-Litterman-Bayes (BLB) model as per Definition 6.1, the associated BLB optimal portfolio is defined to be

$$\boldsymbol{h}^* = \operatorname*{argmax}_{\boldsymbol{h}} \mathbb{E}[u(\boldsymbol{h}'\boldsymbol{r}) \,|\, \boldsymbol{q}]$$

where $\mathbb{E}[\,\cdot\,|\,\boldsymbol{q}]$ denotes the expectation with respect to the posterior predictive density for the random variable $\boldsymbol{r}$. In other words, $\boldsymbol{h}^*$ maximizes posterior expected utility. Explicitly, the posterior predictive density of $\boldsymbol{r}$ is given by

$$\begin{aligned} p(\boldsymbol{r} \,|\, \boldsymbol{q}) &= \int p(\boldsymbol{r} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \boldsymbol{q}) d\boldsymbol{\theta} \quad \text{where} \\ p(\boldsymbol{\theta} \,|\, \boldsymbol{q}) &= \frac{f(\boldsymbol{q} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{q} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \end{aligned}$$

Definition 6.3. Given a benchmark portfolio with holdings $\boldsymbol{h}_B$ (eg. the market portfolio), and given a Black-Litterman-Bayes model (Def. 6.1), the prior $\pi(\boldsymbol{\theta})$ is said to be *benchmark-optimal* if $\boldsymbol{h}_B$ maximizes expected utility of wealth, where the expectation is taken with respect to the *a priori* distribution on asset returns $p(\boldsymbol{r}) = \int p(\boldsymbol{r} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$, so

$$\boldsymbol{h}_B = \operatorname*{argmax}_{\boldsymbol{h}} \int u(\boldsymbol{h}'\boldsymbol{r}) p(\boldsymbol{r} \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{6.16}$$

Many existing approaches are special cases of the above. The model of Black and Litterman (1991) is the special case in which $\boldsymbol{r} \,|\, \boldsymbol{\theta}$ is multivariate normal with mean $\boldsymbol{\theta}$ and $f(\cdot \,|\, \cdot)$ is the normal likelihood for a regression of the portfolio manager's views, the utility of final wealth is any increasing, concave utility function, and the prior is the unique normal distribution which is benchmark-optimal with respect to the market portfolio.

An interesting feature of the model is that there are *two* functions which both play the role of likelihood functions: $p(\boldsymbol{r} \mid \boldsymbol{\theta})$ and $f(\boldsymbol{q} \mid \boldsymbol{\theta})$. Equivalently, we have a triple of random vectors: $(\boldsymbol{r}, \boldsymbol{q}, \boldsymbol{\theta})$ which are not pairwise independent, but $\boldsymbol{r}$ and $\boldsymbol{q}$ are *conditionally independent* given $\boldsymbol{\theta}$. In Bayesian statistics, such situations are commonplace. A *Bayesian network* (or "graphical model") is, intuitively, an arbitrary collection of random variables whose conditional independence structure is specified by a (typically directed and acyclic) graph, so this system could be considered a Bayesian network with three nodes. We refer the reader to Pearl (2014) for the authoritative treatise on Bayesian networks, but suffice it to say that inference with much larger networks than the $(\boldsymbol{r}, \boldsymbol{q}, \boldsymbol{\theta})$ network is now commonplace.

Even if $\boldsymbol{\theta}$ simply represents the mean vector of asset returns, such returns are widely recognized to be non-normal. Replacing (6.4) with a Laplace distribution may fit empirical asset returns more accurately. This corresponds to Lasso regression, a special case of Bayesian regression, in the same sense that the original Black–Litterman model is similar to ridge regression. Giacometti et al. (2007) also investigated heavy-tailed distributions in the context of Black–Litterman optimization.

More generally, $\boldsymbol{\theta}$ is allowed to be any set of parameters appearing in a parametric statistical model for asset returns, not necessarily their means. We explore this class of generalizations in the next sections.

6.4. **APT and Factor Models.** Generalizing further, the parameter vector $\boldsymbol{\theta}$ could represent means (and covariances) of unobservable latent factors in an APT model (Ross, 1976; Roll and Ross, 1980). Such models assume a linear functional form

$$\boldsymbol{r} = \mathbf{X}\boldsymbol{f} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = 0, \quad \mathbb{V}[\boldsymbol{\epsilon}] = \mathbf{D} \tag{6.17}$$

where $\boldsymbol{r}$ is an n-dimensional random vector containing the cross-section of returns in excess of the risk-free rate over some time interval $[t, t+1]$, and $\mathbf{X}$ is a (non-random) $n \times k$ matrix that is known before time t. Also, $\boldsymbol{\epsilon}$ is assumed to follow a mean-zero distribution with diagonal variance-covariance matrix given by

$$\mathbf{D} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) \;\text{ with all }\; \sigma_i^2 > 0. \tag{6.18}$$

The variable $\boldsymbol{f}$ in (6.17) denotes a k-dimensional random vector process which cannot be observed directly; information about the $\boldsymbol{f}$-process must be obtained via statistical inference.

Specifically, we assume that the $\boldsymbol{f}$-process has finite first and second moments given by

$$\mathbb{E}[\boldsymbol{f}] = \boldsymbol{\mu}_f, \quad \text{and} \quad \mathbb{V}[\boldsymbol{f}] = \boldsymbol{F}. \tag{6.19}$$

When necessary, we will use $\boldsymbol{f}_t$ to denote a realization of the $\boldsymbol{f}$-process on day t, but we will usually suppress the implicit time subscript.

The model (6.17), (6.18) and (6.19) entails associated reductions of the first and second moments of the asset returns:

$$\mathbb{E}[\boldsymbol{r}] = \mathbf{X}\boldsymbol{\mu}_f, \quad \text{and} \quad \boldsymbol{\Sigma} := \mathbb{V}[\boldsymbol{r}] = \mathbf{D} + \mathbf{X}\boldsymbol{F}\mathbf{X}' \tag{6.20}$$

where $\mathbf{X}'$ denotes the transpose of $\mathbf{X}$. The elements of $\boldsymbol{\mu}_f$ are called *factor risk premia*. We will continue to use $\boldsymbol{\Sigma}$ to denote $\mathbf{D} + \mathbf{X}\boldsymbol{F}\mathbf{X}'$ throughout this section, and (6.18) implies that $\boldsymbol{\Sigma}^{-1}$ exists.

For simplicity, we treat $\mathbf{X}$ as non-stochastic and assume $k \ll n$. Then (6.20) reduces the number of parameters necessary to describe the density $p(\boldsymbol{r})$ from $O(n^2)$ down to the k parameters in $\boldsymbol{\mu}_f$, the $k(k+1)/2$ parameters in $\boldsymbol{F}$, and n parameters in $\mathbf{D}$, for a total of $n + k(k+3)/2$. Models of the form (6.17) are ubiquitous in practice, and for good reason: in equity markets n is too large to allow direct estimation of $\boldsymbol{\Sigma}$. See Fabozzi, Focardi, and Kolm (2010) and Connor, Goldberg, and Korajczyk (2010) for more discussion and examples.

In the language of Def. 6.1, we are free to choose $\boldsymbol{\theta}$ as any vector of parameters appearing in a parametric statistical model for asset returns; (6.17)-(6.19) is such a model, so as a starting point, choose $\boldsymbol{\theta} = \boldsymbol{\mu}_f$, the k parameters describing the factor risk premia. For simplicity we treat $\boldsymbol{F}$ as a constant matrix, just as the original Black-Litterman model treats $\boldsymbol{\Sigma}$ as a constant matrix.

What kinds of views on factor risk premia do we expect portfolio managers to have? The simplest and most parsimonious scenario is that we have a view on each factor risk premium that is independent of our views on other factors. For example, consider value and momentum, as discussed at length by Asness, Moskowitz, and Pedersen (2013), and Fabozzi, Focardi, and Kolm (2006) and Fabozzi, Focardi, and Kolm (2010) going back to work of Fama and French (1993) and Carhart (1997).

A quantitative portfolio manager might have two views: (1) a view on the value premium, and, separately from that, (2) a view on the momentum premium. It would be atypical for portfolio managers to have views on, say, the sum or difference of the value and momentum premia, or more generally on "portfolios of factors." Hence to keep things simple but still useful, we take the likelihood function to be

$$f(\boldsymbol{q} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{k} \exp[-\frac{1}{2\omega_i^2}(\theta_i - q_i)^2] \tag{6.21}$$

The choice of prior $\pi(\boldsymbol{\theta})$ is very interesting. We discuss two types of priors: one driven by historical data, and one driven by the desire to have some specific benchmark turn out to be optimal under the model of the prior as in in Def. 6.3.

If the random process model driving the unobservable factor returns $\boldsymbol{f}_t$ is stationary, ie. $\boldsymbol{\mu}_f, \boldsymbol{F}$ are approximately constant over time, then we could obtain a prior for $\boldsymbol{\theta} = \boldsymbol{\mu}_f$ by taking the posterior from a simple Bayesian time-series model for the factor returns $\boldsymbol{f}_t$. In particular, the historical mean of the OLS estimates $\hat{\boldsymbol{f}}_t = (\mathbf{X}_t'\mathbf{X}_t)^{-1}\mathbf{X}_t'\boldsymbol{r}_{t+1}$ could be taken as the prior mean. More generally, this problem lends itself well to a hierarchical (or mixed-effects model) approach. Each time period is a "group" and one has models for $\boldsymbol{r}_{t+1} \sim N(\mathbf{X}_t\boldsymbol{f}_t, \mathbf{D})$ and the various $\boldsymbol{f}_t$ are modeled as i.i.d. draws $\boldsymbol{f}_t \sim N(\boldsymbol{\mu}_f, \boldsymbol{F})$. The statistical inference problem is then to infer $\boldsymbol{\theta} = \boldsymbol{\mu}_f$ from observations of $\boldsymbol{r}_t$, a special case of the hierarchical approach discussed in Gelman et al. (2003, Ch. 15). The posterior from this procedure is a possible prior for use in the Black-Litterman procedure.

The "data-driven" approach to prior selection that we have just described has the advantage of not requiring a benchmark portfolio. This makes sense for absolute return strategies where the effective benchmark is cash. It's very common in Bayesian statistics for the posterior from one analysis to become the prior for subsequent analysis.

Alternatively, if there is a benchmark portfolio $\boldsymbol{h}_B$, then closest in spirit to Black and Litterman (1991) would be to search for a benchmark-optimal prior, as defined above. To progress any further, we need to introduce notation for the hyperparameters in $\pi(\boldsymbol{\theta})$, so let's say $\pi(\boldsymbol{\theta}) \sim N(\boldsymbol{\xi}, \boldsymbol{V})$ with $\boldsymbol{\xi} \in \mathbb{R}^k$ and $\boldsymbol{V} \in S^k_{++}$, the set of symmetric positive definite $k \times k$ matrices. Choosing a prior then amounts to choosing $\boldsymbol{\xi}$ and $\boldsymbol{V}$, which are constrained by (6.16). The first step in evaluating (6.16) is to compute the a priori density on returns, $\int p(\boldsymbol{r}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Since $\pi(\boldsymbol{\theta})$ and $p(\boldsymbol{r}\,|\,\boldsymbol{\theta})$ are both Gaussian, this is another completion of squares.

We continue to use the notation $\boldsymbol{\Sigma} = \mathbf{D} + \mathbf{X}\boldsymbol{F}\mathbf{X}'$ as above, since this is the asset-level covariance in an APT model. Straightforward calculations then show:

$$-2\log[p(\boldsymbol{r}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})] = -2\log N(\boldsymbol{r};\ \mathbf{X}\boldsymbol{\theta},\boldsymbol{\Sigma}) - 2\log N(\boldsymbol{\theta};\ \boldsymbol{\xi},\boldsymbol{V})$$

$$\begin{aligned} &= (\boldsymbol{r}-\mathbf{X}\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{r}-\mathbf{X}\boldsymbol{\theta}) + (\boldsymbol{\theta}-\boldsymbol{\xi})'\boldsymbol{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\xi}) \\ &= \boldsymbol{\theta}'\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\eta}'\boldsymbol{\theta} + \boldsymbol{z} \end{aligned}$$

where for notational simplicity, we have introduced the auxiliary variables

$$\boldsymbol{H} = \boldsymbol{V}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X},\ \boldsymbol{\eta} = (\boldsymbol{\xi}'\boldsymbol{V}^{-1} + \boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\mathbf{X})'$$

and

$$\boldsymbol{z} = \boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r} + \boldsymbol{\xi}'\boldsymbol{V}^{-1}\boldsymbol{\xi}.$$

Completing the square again,

$$\boldsymbol{\theta}'\boldsymbol{H}\boldsymbol{\theta} - 2\boldsymbol{\eta}'\boldsymbol{\theta} + \boldsymbol{z} = (\boldsymbol{\theta}-\boldsymbol{v})'\boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{v}) - \boldsymbol{v}'\boldsymbol{H}\boldsymbol{v} + \boldsymbol{z}, \quad \boldsymbol{v} = \boldsymbol{H}^{-1}\boldsymbol{\eta}$$

The integral over $\boldsymbol{\theta}$ is then a standard Gaussian integral, which is performed via the formula

$$\int \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{v})'\boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{v})\right]\,d\boldsymbol{\theta} = \sqrt{\frac{(2\pi)^k}{\det\boldsymbol{H}}}$$

Hence,

$$\int p(\boldsymbol{r}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = (2\pi)^{k/2}\,|\boldsymbol{H}|^{-1}\exp[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\eta}'\boldsymbol{H}^{-1}\boldsymbol{\eta})]$$

$$\begin{aligned} = \frac{(2\pi)^{k/2}}{\det\boldsymbol{H}}\exp\Big[-\frac{1}{2}\Big\{&\boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r} + \boldsymbol{\xi}'\boldsymbol{V}^{-1}\boldsymbol{\xi} \\ &- (\boldsymbol{\xi}'\boldsymbol{V}^{-1} + \boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\mathbf{X})\boldsymbol{H}^{-1}(\boldsymbol{V}^{-1}\boldsymbol{\xi} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r})\Big\}\Big] \end{aligned}$$

Let's multiply out the second quadratic term:

$$\begin{aligned} &(\boldsymbol{\xi}'\boldsymbol{V}^{-1} + \boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\mathbf{X})\boldsymbol{H}^{-1}(\boldsymbol{V}^{-1}\boldsymbol{\xi} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r}) \\ &= \boldsymbol{\xi}'(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V})^{-1}\boldsymbol{\xi} + 2\boldsymbol{\xi}'(\boldsymbol{H}\boldsymbol{V})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r} \\ &\quad + \boldsymbol{r}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{r} \end{aligned}$$

Note that $\int p(\boldsymbol{r}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is a Gaussian probability distribution for the random vector $\boldsymbol{r}$, so to find the covariance, we just collect the quadratic terms in $\boldsymbol{r}$ and take the inverse:

$$\mathbb{V}_\pi[\boldsymbol{r}] = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})^{-1}.$$

Completing the squares as before, the mean is

$$\begin{aligned}\mathbb{E}_\pi[\boldsymbol{r}] &= \mathbb{V}_\pi[\boldsymbol{r}]\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\xi} \\ &= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\xi}\end{aligned}$$

The *a priori* optimal portfolio under CARA utility is of course

$$(\delta\mathbb{V}_\pi[\boldsymbol{r}])^{-1}\mathbb{E}_\pi[\boldsymbol{r}] = \delta^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{H}^{-1}\boldsymbol{V}^{-1}\boldsymbol{\xi} \tag{6.22}$$

but unlike the classic Black-Litterman case, it is no longer true that any arbitrary benchmark portfolio can be realized as an *a priori* optimal portfolio. In fact, (6.22) gives a very simple characterization of those that can: they are precisely of the form $\delta^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{\Pi}$ where $\mathbf{\Pi}$ is some linear combination of the columns of $\mathbf{X}$. That is to say, they are portfolios which are optimal with respect to a set of individual asset risk premia that come from the factor model. From the standpoint of APT, this is not a real restriction; if the original APT model isn't mis-specified, then residuals should be independent, and not additional sources of risk premia for use in forming expected returns.

Not every possible portfolio is realizable as *a priori* optimal, hence the market portfolio may not be. However, at least we can say that if the market is in a CAPM equilibrium and if one of the columns of $\mathbf{X}$ contains the CAPM betas, then the individual asset risk premia will be proportional to that column of $\mathbf{X}$, and then the market portfolio *will* be realizable as *a priori* optimal, as per (6.22).

We now leave behind the question of the prior and continue with calculating the *a posteriori* optimal portfolio, i.e. the portfolio which takes into account the views (6.21) on the factor risk premia. This calculation proceeds in three steps:

1. calculate the posterior distribution of $\boldsymbol{\theta}$, after the views are taken into account, which is given by

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{q}) = \frac{f(\boldsymbol{q}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{q}\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

2. calculate the *a posteriori* distribution of asset returns (also called the posterior predictive density), given by
$$p(\boldsymbol{r} \mid \boldsymbol{q}) = \int p(\boldsymbol{r} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{q}) d\boldsymbol{\theta} \tag{6.23}$$
3. calculate the mean-variance optimal portfolio under (6.23).

Fortunately, Step 1 is easy since the normal prior is a *conjugate prior* for the normal likelihood, meaning that the posterior distribution is of the same distributional family as the prior (again normal), but with different values for the hyperparameters. By a straightforward calculation, if the prior is normal with hyperparameters $\boldsymbol{\xi}, \boldsymbol{V}$ then the posterior has hyperparameters $\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{V}}$ where
$$\tilde{\boldsymbol{V}} = (\boldsymbol{V}^{-1} + \boldsymbol{\Omega}^{-1})^{-1}, \quad \tilde{\boldsymbol{\xi}} = (\boldsymbol{V}^{-1} + \boldsymbol{\Omega}^{-1})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{\xi} + \boldsymbol{\Omega}^{-1}\boldsymbol{q})$$

Step 2 follows via the same calculation we did to find the *a priori* density, but using the posterior updated values $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\xi}}$ for the hyperparameters. We don't need to do the whole calculation again, just make the substitution $\boldsymbol{\xi} \to \tilde{\boldsymbol{\xi}}$ and $\boldsymbol{V} \to \tilde{\boldsymbol{V}}$ to find
$$\mathbb{V}[\boldsymbol{r} \mid \boldsymbol{q}] = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\tilde{\boldsymbol{V}}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})^{-1}.$$
$$\mathbb{E}[\boldsymbol{r} \mid \boldsymbol{q}] = \mathbb{V}[\boldsymbol{r} \mid \boldsymbol{q}]\boldsymbol{\Sigma}^{-1}\mathbf{X}(\tilde{\boldsymbol{V}}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\xi}} \tag{6.24}$$

Step 3 is then a completely standard calculation of a mean-variance optimal portfolio from (6.24):
$$\begin{aligned} \boldsymbol{h}^* &= \delta^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi} \qquad (6.25) \\ \boldsymbol{\Pi} &:= \mathbf{X}\tilde{\boldsymbol{\mu}}_f \\ \tilde{\boldsymbol{\mu}}_f &:= (\boldsymbol{V}^{-1} + \boldsymbol{\Omega}^{-1} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{\xi} + \boldsymbol{\Omega}^{-1}\boldsymbol{q}) \end{aligned}$$

Eqns. (6.25) represent the solution to Black-Litterman optimization in the context of APT. They are written in a suggestive form: the asset-level risk premia $\boldsymbol{\Pi} = \mathbf{X}\tilde{\boldsymbol{\mu}}_f$ are linear combinations of the factors which form the columns of $\mathbf{X}$. One can think of $\tilde{\boldsymbol{\mu}}_f$ as a set of factor risk premia "adjusted" to take account of the views, and the adjustments tend to give more weight to factors which have high prior mean-variance ratios ξ_i/V_{ii} and/or high expected return-uncertainty ratios Q_i/ω_i^2.

6.5. **Concluding Remarks.** Many practical problems now require inference in "data scarce" situations where the number of parameters may greatly exceed the number of observations. Black-Litterman optimization with a small number of views is one such problem.

Any Bayesian statistical model of asset returns, together with a utility function of final wealth in the sense of Arrow (1971) and Pratt (1964), gives rise to an associated Black-Litterman-Bayes optimization procedure. Key to the generality of this procedure is that $\boldsymbol{\theta}$ can be any vector of parameters appearing in a statistical model for asset returns, and need not be simply a parameter representing the mean return. Section 6.3 also shows that our Black-Litterman-Bayes generalization is itself a special case of a Bayesian network in the sense of Pearl (2014).

In the process we hope to have clarified the precise sense in which the original model of Black and Litterman (1991) and Black and Litterman (1992) is "Bayesian." Usually, in Bayesian statistics, the likelihood function plays the essential role of connecting the empirical data to the parameters. The structure of the experimental design must be encoded in the likelihood function, which is why the matrix in a regression is called the "design matrix."

In Black–Litterman type models, the likelihood isn't really what connects the parameters to empirical data, unless we broaden the definition of "empirical data" to include data on portfolio managers' views. We reconcile this by regarding portfolio managers' views as (possibly *very* noisy) observations of the future, and as such, operationally no different from empirical data obtained through a very unreliable measuring device.

References

Arrow, Kenneth J (1971). "Essays in the theory of risk-bearing". In:

Asness, Clifford S, Tobias J Moskowitz, and Lasse Heje Pedersen (2013). "Value and momentum everywhere". In: *The Journal of Finance* 68.3, pp. 929–985.

Black, Fischer and Robert Litterman (1992). "Global portfolio optimization". In: *Financial Analysts Journal*, pp. 28–43.

Black, Fischer and Robert B Litterman (1991). "Asset allocation: combining investor views with market equilibrium". In: *The Journal of Fixed Income* 1.2, pp. 7–18.

Carhart, Mark M (1997). "On persistence in mutual fund performance". In: *The Journal of Finance* 52.1, pp. 57–82.

Connor, Gregory, Lisa R Goldberg, and Robert A Korajczyk (2010). *Portfolio risk analysis*. Princeton University Press.

Fabozzi, Frank J, Sergio M Focardi, and Petter N Kolm (2006). "Incorporating trading strategies in the Black-Litterman framework". In: *The Journal of Trading* 1.2, pp. 28–37.

– (2010). *Quantitative equity investing: Techniques and Strategies*. John Wiley & Sons.

Fama, Eugene F and Kenneth R French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56.

Gelman, Andrew et al. (2003). *Bayesian data analysis, 2nd ed.* Taylor & Francis.

Giacometti, Rosella et al. (2007). "Stable distributions in the Black–Litterman approach to asset allocation". In: *Quantitative Finance* 7.4, pp. 423–433.

He, Guangliang and Robert Litterman (1999). "The intuition behind Black-Litterman model portfolios". In: *Goldman Sachs Investment Management Series*.

Pearl, Judea (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pratt, John W (1964). "Risk aversion in the small and in the large". In: *Econometrica: Journal of the Econometric Society*, pp. 122–136.

Robert, Christian (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Roll, Richard and Stephen A Ross (1980). "An empirical investigation of the arbitrage pricing theory". In: *The Journal of Finance* 35.5, pp. 1073–1103.

Ross, Stephen A (1976). "The arbitrage theory of capital asset pricing". In: *Journal of economic theory* 13.3, pp. 341–360.

Satchell, Stephen and Alan Scowcroft (2000). "A demystification of the Black–Litterman model: Managing quantitative and traditional portfolio construction". In: *Journal of Asset Management* 1.2, pp. 138–150.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.