7. Black-Litterman Optimization

7.1. **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,

$$\begin{aligned} p(A \,|\, B)p(B) &= p(A \cap B) = p(B \cap A) = p(B \,|\, A)p(A) \\ p(A \,|\, B) &= p(B \,|\, A)\frac{p(A)}{p(B)}\,. \end{aligned} \tag{7.1}$$

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

Eq. (7.1) 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.

The following is standard terminology in the statistics literature (Robert, 2007). A *Bayesian statistical model* consists of:

(1) a vector-valued random variable $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ distributed according to $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$, where realizations of $\boldsymbol{x}$ 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 $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ 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*

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

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

Example 7.1 (Bayes (1763)). A ball A is rolled along the unit interval $[0,1]$, with a uniform probability of stopping at any point. It stops at a point θ between 0 and 1, and is not moved subsequently. A second ball B is then rolled n times under the same assumptions. Let x denote the number of times the ball B stopped before passing A. Given x, what inference can we make on θ?

Once θ is known, stopping to the left versus the right of θ is just like flipping an unfair coin with prob(heads) $= \theta$. Experiments like flipping an unfair coin are known in mathematics as *Bernoulli trials*.

Formally, a *Bernoulli trial* is any random experiment with only two outcomes, usually called "success" and "failure" and conveniently denoted by 1 or 0. A *Bernoulli process* is a finite or infinite sequence of independent Bernoulli trials with the same probability of success, ie. a sequence of independent random variables $X_1, X_2, X_3, \ldots$ with $X_i \in \{0,1\}$, and there exists some $\theta \in [0,1]$ such that $P(X_i = 1) = \theta$ for all $i \in \mathbb{N}$. The number of successes in n independent Bernoulli trials is a random variable whose distribution is called the *Binomial distribution* and denoted $\mathrm{B}(n,\theta)$.

Therefore, in Example 7.1, $x \sim \mathrm{B}(n,\theta)$. Bayes' theorem tells us

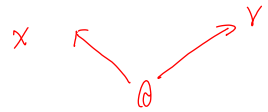
$$p(\theta \,|\, x) \propto p(x \,|\, \theta) p(\theta). \tag{7.2}$$

By assumption θ is uniform on $[0,1]$, so $p(\theta) = 1$. Hence (7.2) is proportional to the binomial density $p(x \,|\, \theta) \propto \theta^x (1-\theta)^{n-x}$ considered as a function of θ, which is known as the *beta distribution* $\mathrm{Beta}(x+1, n-x+1)$.

This example illustrates that the prior isn't necessarily something connected with opinions or views. Sometimes it is dictated by physics. In this example, the uniform prior is determined by the mechanics of the process which occurred first (the rolling of the first ball).

7.2. **Prediction.** 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 $\boldsymbol{r}$:

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



Let's consider prediction from the unfair-coin model. What is the probability that the next flip, say y, is heads (denoted by 1 for convenience).

$$p(y = 1 \,|\, x) = \int p(y = 1 \,|\, \theta, x) p(\theta \,|\, x) \, d\theta = \frac{x+1}{n+2}.$$

This is known as "Laplace's law of succession." Note that

$$p(y = 1 \,|\, \theta, x) = \theta,$$

so our forecast is just the posterior mean. The mean of Beta(α, β) is $\alpha/(\alpha+\beta)$.

7.3. **Conjugate priors.** Now suppose that, instead of the uniform prior implied by Bayes' example, we take Beta(α, β) as our prior. In other words, we take as prior the following:

$$p(\theta \,|\, \alpha, \beta) \;\propto\; \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The posterior is then

$$\begin{aligned} p(\theta \,|\, x) &\propto p(x \,|\, \theta) p(\theta) \\ &\propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\sim \text{Beta}(\alpha + x, \beta + n - x). \end{aligned}$$

也就是说 liklihood 都分是 binomial, prior 和 posterior 的从都是 beta distribution，这叫 conjugate.

Thus the posterior is of the same form as the prior, but with different parameters. The beta distribution is said to be a *conjugate prior* for the binomial model.

We also say the prior has been "updated" by the data, x in this case. If more data come in, the posterior is suitable for use as a prior in the next round of updating. The order in which i.i.d. observations are collected does not matter, and updating the prior one observation at a time, or all observations together, does not matter. This kind of model lends itself well to inference with big data sets, since all of the relevant information about the full data set – no matter how large it may be – is contained in a small number of parameters.

7.4. **The Normal Distribution.** Old friends are the best kind. Specifically, consider an i.i.d. sample $x = (x_1, \ldots, x_n)$ drawn from a normal distribution $N(\mu, \sigma^2)$. It is convenient to write $\tau = 1/\sigma^2$. In this course we will generally always use θ to denote the vector of parameters in the model. In this case $\theta = (\mu, \tau)$.

Consider the prior

$$\tau \sim \Gamma(\alpha, \beta), \quad \mu \,|\, \tau \sim N(\nu, \frac{1}{k\tau}) \;\text{ for } k > 0, \nu \in \mathbb{R}.$$

This prior distribution for τ involves the Gamma distribution, which exists in several different parameterizations. The one we'll use has pdf given by

$$\beta^{\alpha} x^{\alpha-1} e^{-x\beta}/\Gamma(\alpha) \text{ for } x \geq 0 \text{ and } \alpha, \beta > 0.$$

In this case α is called the shape parameter, while β is called the inverse-scale or rate parameter.

If $\tau \sim \Gamma(\alpha, \beta)$ with these conventions, this means that $\sigma^2 \sim \text{IG}(\alpha, \beta)$ follows the aptly-named inverse-gamma distribution. The inverse gamma distribution's pdf is defined over the support $x > 0$ and is given by

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right).$$

Note that to get the density of σ^2 from the density of $\tau = 1/\sigma^2$ one would use the change-of-variables formula. Aside: This distribution is not only useful as a Bayesian prior; the first hitting time of a standard Wiener process has an inverse-gamma distribution with parameters $\alpha = 1/2$ and $\beta = x^2/2$ where x is the value to hit.

The full prior density is

$$\pi(\tau, \mu) = \pi(\tau)\pi(\mu \,|\, \tau),$$

which may be written as

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \cdot \sqrt{\frac{k\tau}{2\pi}} \exp\Big\{ -\frac{k\tau}{2}(\mu - \nu)^2 \Big\}.$$

or more simply

$$\pi(\tau, \mu) \propto \tau^{\alpha-1/2} \exp\left[-\tau \Big\{ \beta + \frac{k}{2}(\mu - \nu)^2 \Big\} \right].$$

We have $x_1, \ldots, x_n$ independent, identically distributed from $N(\mu, 1/\tau)$, so the likelihood is

$$p(x \,|\, \mu, \tau) = \Big(\frac{\tau}{2\pi}\Big)^{n/2} \exp\Big\{ -\frac{\tau}{2} \sum_i (x_i - \mu)^2 \Big\}$$

Thus

$$\pi(\tau, \mu \,|\, x) \propto \tau^{\alpha+n/2-1/2} \exp\left[-\tau \Big\{ \beta + \frac{k}{2}(\mu - \nu)^2 + \frac{1}{2} \sum_i (x_i - \mu)^2 \Big\} \right].$$

Complete the square to see that

$$k(\mu - \nu)^2 + \sum_i (x_i - \mu)^2 = (k + n) \left(\mu - \frac{k\nu + n\bar{x}}{k + n} \right)^2 + \frac{nk}{n + k} (\bar{x} - \nu)^2 + \sum_i (x_i - \bar{x})^2.$$

Hence the posterior satisfies

$$\pi(\tau, \mu \,|\, x) \propto \tau^{\alpha'-1/2} \exp\left[-\tau \Big\{ \beta' + \frac{k'}{2}(\mu - \nu')^2 \Big\} \right]$$

where

$$\begin{aligned} \alpha' &= \alpha + \frac{n}{2} \\ \beta' &= \beta + \frac{1}{2} \frac{nk}{n + k} (\bar{x} - \nu)^2 + \frac{1}{2} \sum_i (x_i - \bar{x})^2 \\ k' &= k + n \\ \nu' &= \frac{k\nu + n\bar{x}}{k + n}. \end{aligned}$$

$$= \frac{k}{k+n} \nu + \frac{n}{k+n} \bar{x}$$

原Assumption中 $\mu|\tau \sim N(\nu, \frac{1}{k\tau})$

由于加入新数据 $\mu|\tau$ 它的原方差由 $\frac{\sigma^2}{k} \to \frac{\sigma^2}{n+k}$，变小，

原均值变为 ν 与 $\bar{x}$ 的 weighted average

ν: Assumption 均值

$\bar{x}$: 样本均值

样本数量

极点关于 μ 的二次项
关于 μ 的一次项
和常数项

$a(u-b)^2 = au^2 - 2abu + ab^2$
u^2 的系数就为 a
b 可由 u 的系数 $\div -2a$ 得到
再求常数项.

Thus the posterior distribution is of the same parametric form as the prior (the above form of prior is a conjugate family), but with (α, β, k, ν) replaced by $(\alpha', \beta', k', \nu')$. To write code to update this model with new data, all you really need is to keep track of the sufficient statistics! Equivalently you could keep track of the parameters (α, β, k, ν).

7.5. **Bayesian Multivariate Regression.** We now proceed with a fully Bayesian multivariate analysis of the likelihood

$$p(y \,|\, \theta) = (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma^2} \|y - X\beta\|^2 \Big] \tag{7.3}$$

After completing the square, (7.3) can be written as follows:

$$\ell(\beta, \sigma^2 \,|\, y) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \Big\{ s^2 + (\beta - \hat\beta)'(X'X)(\beta - \hat\beta) \Big\} \right] \tag{7.4}$$

Eq. (7.4) implies that $T(y) = (s^2, \hat\beta)$ forms a $(k+1)$-dimensional (minimal) sufficient statistic for the parameters $\theta = (\sigma^2, \beta)$.

In real-world problems with large data sets, we are concerned with "online" updating. In an online scenario, we are concerned with how we will update the sufficient statistics if a new row of X and a new response variable are published. The answer to this question is that we can update the $k + k(k+1)/2$ numbers $X'X$ and $X'y$, and use those to update the minimal sufficient statistics $(\hat\beta, s^2)$.

A conjugate prior can be constructed within the Gaussian-Inverse-Gamma family:

$$\text{prior:} \quad \beta \,|\, \sigma^2 \sim N(\beta_0, \sigma^2 M^{-1}), \quad \sigma^2 \sim \text{IG}(a, b), \quad a, b > 0. \tag{7.5}$$

The joint posterior density of β, σ^2 is

$$p(\beta, \sigma^2 \,|\, y) = p(\beta \,|\, y, \sigma^2) p(\sigma^2 \,|\, y)$$

so it suffices to compute $p(\beta \,|\, y, \sigma^2)$ and $p(\sigma^2 \,|\, y)$.

In order to express the various conditional and marginal posteriors, we define a few variables to simplify notation:

$$\Omega := (M + X'X)^{-1} \tag{7.6}$$

$$\hat\mu = (M + X'X)^{-1}(X'X\hat\beta + M\beta_0) \tag{7.7}$$

$$Q := (\beta_0 - \hat\beta)'(M^{-1} + (X'X)^{-1})^{-1}(\beta_0 - \hat\beta) \tag{7.8}$$

Then the conditional posterior $p(\beta \,|\, \sigma^2, y)$ is normal with

$$\beta \,|\, \sigma^2, y \sim N(\hat\mu,\ \sigma^2\Omega).$$

Note that if $M = \lambda I$ and $\beta_0 = 0$, then $\hat\mu$ is the ridge regression estimator $\hat\beta_\lambda$.

The marginal posterior for σ^2 is:

$$\sigma^2 \,|\, y \sim \mathrm{IG}\Big(\frac{n}{2} + a,\ \frac{2b + s^2 + Q}{2}\Big) \tag{7.9}$$

Note that the posteriors for β and σ^2 are of the same types as the prior (7.5), i.e. (7.5) is a conjugate prior.

Integrating out σ^2 from $p(\beta \,|\, \sigma^2, y)$ using the marginal posterior (7.9) yields the marginal posterior of β:

$$\beta \,|\, y \sim \mathcal{T}_{k+1}\left(n + 2a, \hat{\mu}, \frac{2b + s^2 + Q}{n + 2a}\Omega\right).$$

where the density of $\mathcal{T}_p(\nu, \theta, \Sigma)$ is defined as

$$f(t \,|\, \nu, \theta, \Sigma) = \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2)\sqrt{|\Sigma|\nu\pi}}\left[1 + \frac{(t - \theta)'\Sigma^{-1}(t - \theta)}{\nu}\right]^{-(\nu+p)/2}.$$

Hence $\hat{\mu} = \mathbb{E}[\beta \,|\, y]$ is the posterior mean, which is the Bayes estimator associated to quadratic loss; it can be viewed as a weighted average between the maximum likelihood estimator and the prior mean.

7.6. **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 Litterman and He (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 Litterman and He (1999). Let $\boldsymbol{p} \in \mathbb{R}^n$ denote a dollar-neutral 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

$$\mathbb{E}[\boldsymbol{p}'\boldsymbol{r}] = q \in \mathbb{R} \tag{7.10}$$

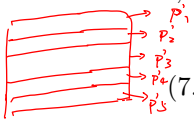
where $\boldsymbol{r}$ is the random vector of asset returns over some subsequent interval, and q denotes the expected return, according to the view.

Suppose that the total number of such views is k, where each individual view is analogous to the above but with a different portfolio and corresponding expected return:

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

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

$$\mathbb{E}[\boldsymbol{P}\boldsymbol{r}] = \boldsymbol{q} \text{ for } \boldsymbol{q} \in \mathbb{R}^k. \tag{7.11}$$



The core idea of Black and Litterman (1991) is to treat the portfolio manager's views (7.11) as noisy observations which are useful for performing statistical inference concerning the parameters in some underlying model for $\boldsymbol{r}$. For example, if

$$\boldsymbol{r} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{7.12}$$

with $\boldsymbol{\Sigma}$ a known positive-definite $n \times n$ matrix, then the views (7.11) 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 some underlying model for asset returns, such as (7.12). This is expressed as the following (more precise) restatement of (7.11):

$$\boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{q} + \boldsymbol{\epsilon}^{(v)}, \quad \boldsymbol{\epsilon}^{(v)} \sim N(0, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \operatorname{diag}(\omega_1, \dots, \omega_k) \tag{7.13}$$

Eq. (7.13) specifies a Gaussian distribution and may be alternately written in the form of a likelihood:

$$f(\boldsymbol{q} \,|\, \boldsymbol{\theta}) \propto \exp\big[-\frac{1}{2}(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q})'\boldsymbol{\Omega}^{-1}(\boldsymbol{P}\boldsymbol{\theta} - \boldsymbol{q})\big] \tag{7.14}$$

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 (7.13) specifies the likelihood as

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

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 multi-parameter 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

$$\boldsymbol{r} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \quad \text{and} \quad \boldsymbol{\theta} \sim N(\boldsymbol{\Pi}, \boldsymbol{C}) \tag{7.15}$$

for some covariance $\boldsymbol{C}$ whose inverse represents the amount of precision in the prior. For any portfolio $\boldsymbol{p}$, then, according to (7.15) we have

$$\mathbb{E}[\boldsymbol{p}'\boldsymbol{r}] = \boldsymbol{p}'\boldsymbol{\Pi} \quad \mathbb{V}[\boldsymbol{p}'\boldsymbol{r}] = \boldsymbol{p}'(\boldsymbol{\Sigma} + \boldsymbol{C})\boldsymbol{p}\,. \tag{7.16}$$

In fact we must make a choice whether to use the conditional or unconditional variance in optimization:

$$\mathbb{V}(\boldsymbol{r} \,|\, \boldsymbol{\theta}) = \boldsymbol{\Sigma} \;\text{ but }\; \mathbb{V}(\boldsymbol{r}) = \boldsymbol{\Sigma} + \boldsymbol{C}\,.$$

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

$$\boldsymbol{h}_{\text{eq}} = \kappa^{-1}(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1}\boldsymbol{\Pi}. \tag{7.17}$$

Any combination of $\boldsymbol{\Pi}, \boldsymbol{C}$ satisfying (7.17) 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

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

We thus have the normal likelihood (7.14) and the normal prior (7.15) 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). Let $N(x; \mu, \sigma^2)$ denote the probability density of the normal, with μ, σ as parameters and evaluated as a function of x for the density. Up to normalization factors,

$$\text{posterior} \;\propto\; \underbrace{N(\boldsymbol{q}; \boldsymbol{P}\boldsymbol{\theta}, \boldsymbol{\Omega})}_{\text{likelihood}} \times \underbrace{N(\boldsymbol{\theta}; \boldsymbol{\Pi}, \boldsymbol{C})}_{\text{prior}}$$

The normal density is of the form $\exp[-\frac{1}{2}\text{quadratic}]$ so this will be easier to analyze if we take $-2\log$ of both sides.

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

$$(\boldsymbol{P\theta}-\boldsymbol{q})'\boldsymbol{\Omega}^{-1}(\boldsymbol{P\theta}-\boldsymbol{q})+(\boldsymbol{\theta}-\boldsymbol{\Pi})'\boldsymbol{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\Pi}) \tag{7.18}$$

$$=\boldsymbol{\theta}'\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P\theta}-\boldsymbol{\theta}'\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q}-\boldsymbol{q}'\boldsymbol{\Omega}^{-1}\boldsymbol{P\theta} \tag{7.19}$$

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

$$=\boldsymbol{\theta}'[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{C}^{-1}]\boldsymbol{\theta}-2(\boldsymbol{q}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{\Pi}'\boldsymbol{C}^{-1})\boldsymbol{\theta} \tag{7.20}$$

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 $p(\boldsymbol{\theta})$,

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

For the quadratic term to match (7.20) we must have $\boldsymbol{H}=\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{C}^{-1}$ and hence the posterior has mean

$$\boldsymbol{v}=[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{C}^{-1}]^{-1}[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q}+\boldsymbol{C}^{-1}\boldsymbol{\Pi}] \tag{7.21}$$

and covariance

$$\boldsymbol{H}^{-1}=[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{C}^{-1}]^{-1}. \tag{7.22}$$

Part of the beauty of this derivation is its simplicity: going from (7.18) to (7.22) requires just a few lines of algebra. Note that (7.22) is the posterior covariance of $\boldsymbol{\theta}$, not the *a posteriori* covariance of $\boldsymbol{r}$.

If asset returns are modeled using an elliptical distribution (and they are in this case), investors with any concave, increasing utility function will want to solve

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

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

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

Just for fun, let's verify that this approaches $\boldsymbol{h}_{eq}$ when there aren't any views or, equivalently, when $\boldsymbol{\Omega}\to\infty$ or $\boldsymbol{\Omega}^{-1}\to 0$ and $\boldsymbol{C}=\tau\boldsymbol{\Sigma}$:

$$\boldsymbol{H}^{-1}=[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{P}+\boldsymbol{C}^{-1}]^{-1}=\boldsymbol{C}=\tau\boldsymbol{\Sigma}$$

In this case

$$\begin{aligned}\kappa^{-1}[\boldsymbol{H}^{-1}+\boldsymbol{\Sigma}]^{-1}\boldsymbol{H}^{-1}[\boldsymbol{P}'\boldsymbol{\Omega}^{-1}\boldsymbol{q}+\boldsymbol{C}^{-1}\boldsymbol{\Pi}] \\ =[\kappa(1+\tau)\boldsymbol{\Sigma}]^{-1}\tau\boldsymbol{\Sigma}[(\tau\boldsymbol{\Sigma})^{-1}\boldsymbol{\Pi}] \\ =[\kappa(1+\tau)\boldsymbol{\Sigma}]^{-1}\boldsymbol{\Pi}=\boldsymbol{h}_{eq}\end{aligned}$$

which verifies exactly the property we wanted to check.

References

Bayes, Thomas (1763). "An essay towards solving a problem in the doctrine of chances." In:

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.

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

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

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.