Black-Litterman-Bayes, Kalman Filter, ICA

Kaiwen Zhou, Youran Pan, Erding Liao

Contents

1 Topic: Black-Litterman-Bayes 2

2 Topic: Important Properties of the Kalman Filter 4

3 Topic: Kalman filter's Application on Financial Problems — Pairs Trading 7

4 Topic: Kalman filter's Application on Financial Problems — Index Tracking Portfolios 8

5 Topic: Another Interpretation for Independent Component Analysis (ICA) 9

Useful Theorem and Lemma

Theorem 0.1. ***(Binomial Inverse Theorem - Woodbury Matrix Identity)*** *If* $\mathbf{A}$, $\mathbf{U}$, $\mathbf{B}$, $\mathbf{V}$ *are matrices of sizes* $n \times n, n \times k, k \times k, k \times n$*, respectively, then*

(a) If $\mathbf{A}$ *and* $\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}$ *are nonsingular:*

$$(\mathbf{A}+\mathbf{U B V})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1} \mathbf{U B}\left(\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}\right)^{-1} \mathbf{B V A}^{-1}$$

(b) And (a) can be simplified to

$$(\mathbf{A}+\mathbf{U B V})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1} \mathbf{U}\left(\mathbf{B}^{-1}+\mathbf{V A}^{-1} \mathbf{U}\right)^{-1} \mathbf{V A}^{-1}.$$

Solution:

(a) Prove by verification. First notice that

$$(\mathbf{A}+\mathbf{U B V}) \mathbf{A}^{-1} \mathbf{U B}=\mathbf{U B}+\mathbf{U B V A}^{-1} \mathbf{U B}=\mathbf{U}\left(\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}\right).$$

Now multiply the matrix we wish to invert by its alleged inverse

$$\begin{aligned}
&(\mathbf{A}+\mathbf{U B V})\left(\mathbf{A}^{-1}-\mathbf{A}^{-1} \mathbf{U B}\left(\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}\right)^{-1} \mathbf{B V A}^{-1}\right) \\
&=\mathbf{I}_n+\mathbf{U B V A}^{-1}-\mathbf{U}\left(\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}\right)\left(\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}\right)^{-1} \mathbf{B V A}^{-1} \\
&=\mathbf{I}_n+\mathbf{U B V A}^{-1}-\mathbf{U B V A}^{-1} \\
&=\mathbf{I}_n
\end{aligned}$$

which verifies that it is the inverse.

(b) Since $\mathbf{B}+\mathbf{B V A}^{-1} \mathbf{U B}=\mathbf{B}\left(\mathbf{I}+\mathbf{V A}^{-1} \mathbf{U B}\right)$ is nonsingular, we must have $\mathbf{B}$ is invertible. Then the two B terms flanking the quantity inverse in the right-hand side can be replaced with $\left(\mathbf{B}^{-1}\right)^{-1}$, which simplifies (a) to

$$(\mathbf{A}+\mathbf{U B V})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1} \mathbf{U}\left(\mathbf{B}^{-1}+\mathbf{V A}^{-1} \mathbf{U}\right)^{-1} \mathbf{V A}^{-1}$$

Lemma 0.2. *If a multivariate normal random variable* $\boldsymbol{\theta}$ *has density* $p(\boldsymbol{\theta})$ *and*

$$-2 \log p(\boldsymbol{\theta})=\boldsymbol{\theta}^{\top} \boldsymbol{H} \boldsymbol{\theta}-2 \boldsymbol{\eta}^{\top} \boldsymbol{\theta}+(\textit{terms without } \boldsymbol{\theta})$$

then $\mathbb{V}[\boldsymbol{\theta}]=\boldsymbol{H}^{-1}$ *and* $\mathbb{E}[\boldsymbol{\theta}]=\boldsymbol{H}^{-1} \boldsymbol{\eta}$.

Solution: For $\boldsymbol{H}$ symmetric, we have

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

Set $\boldsymbol{v}=\boldsymbol{H}^{-1} \boldsymbol{\eta}$ in the above equation, we obtain

$$-2 \log p(\boldsymbol{\theta})=\boldsymbol{\theta}^{\top} \boldsymbol{H} \boldsymbol{\theta}-2 \boldsymbol{\eta}^{\top} \boldsymbol{\theta}+(\text{terms without } \boldsymbol{\theta})=\left(\boldsymbol{\theta}-\boldsymbol{H}^{-1} \boldsymbol{\eta}\right)^{\top} \boldsymbol{H}\left(\boldsymbol{\theta}-\boldsymbol{H}^{-1} \boldsymbol{\eta}\right)+(\text{ terms without } \boldsymbol{\theta})$$

Therefore, we must have $\mathbb{V}[\boldsymbol{\theta}]=\boldsymbol{H}^{-1}$ and $\mathbb{E}[\boldsymbol{\theta}]=\boldsymbol{H}^{-1} \boldsymbol{\eta}$.

Definition 0.3. ***(Prior Predictive Distribution)*** *The prior predictive distribution is for predicting distribution for* x *BEFORE any sample of* x *has been gathered/observed. The only information we have at this stage is our belief about the prior,* $\pi(\boldsymbol{\theta})$*, and sampling distribution i.e.* $p(x \mid \boldsymbol{\theta})$*. Then, the prior predictive distribution is given by*

$$p(\boldsymbol{x})=\int_{\Theta} p(\boldsymbol{x}, \boldsymbol{\theta}) d \boldsymbol{\theta}=\int_{\Theta} p(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d \boldsymbol{\theta}$$

After the sample has been gathered, we obtain new information, i.e., the likelihood. Then, we can derive the posterior distribution $\theta|x_{observed}$ and use that to predict new values/distribution for x which is given by the posterior predictive distribution:

$$p(\boldsymbol{x}_{\text{ new }} | \boldsymbol{x}_{\text{ observed}}) = \int_\Theta p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{x}_{\text{ observed}}) d\boldsymbol{\theta} = \int_\Theta p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{x}_{\text{ observed}}) p(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{ observed}}) d\boldsymbol{\theta}$$

Lemma 0.4. *If X, Y, W, and V are real-valued random variables and a, b, c, d are real-valued constants, then the following facts are a consequence of the definition of covariance:*

$$\begin{aligned}
\operatorname{cov}(X, a) &= 0 \\
\operatorname{cov}(X, X) &= \operatorname{var}(X) \\
\operatorname{cov}(X, Y) &= \operatorname{cov}(Y, X) \\
\operatorname{cov}(aX, bY) &= ab \operatorname{cov}(X, Y) \\
\operatorname{cov}(X + a, Y + b) &= \operatorname{cov}(X, Y) \\
\operatorname{cov}(aX + bY, cW + dV) &= ac \operatorname{cov}(X, W) + ad \operatorname{cov}(X, V) + bc \operatorname{cov}(Y, W) + bd \operatorname{cov}(Y, V)
\end{aligned}$$

1 Topic: Black-Litterman-Bayes

Problem 1.1. *This question refers to the article "On The Bayesian Interpretation Of Black-Litterman" [1].*

(a) Derive formulas (10)-(11) using the properties of the multivariate normal distribution in the slides "Bayesian Modeling: Introduction".

*(b) **(Extra credit)** Derive formulas (22)-(26) using the same properties.*

Classic Black-Litterman Model from Bayesian Perspective:

Suppose we have $\boldsymbol{r} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, and since 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 Π, the investor's model of the world is that

$$\boldsymbol{r} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\Pi}, \boldsymbol{C}) \quad \text{where } \boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{\Pi} \in \mathbb{R}^n, \boldsymbol{\Sigma}, \boldsymbol{C} \in \mathbb{R}^{n \times n} \tag{1}$$

A key aspect of the model is that 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 $\boldsymbol{\Sigma}$. This is expressed as the following:

$$\boldsymbol{P\theta} = \boldsymbol{q} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \operatorname{diag}(\omega_1, \ldots, \omega_k), \omega_i > 0 \tag{2}$$

where $\boldsymbol{P} \in \mathbb{R}^{k \times n}$, $\boldsymbol{\Omega} \in \mathbb{R}^{k \times k}$ and $\boldsymbol{q}, \boldsymbol{\epsilon} \in \mathbb{R}^k$

(a) **Solution:** In this question, we derive the mean $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{H}$ for the posterior.

Since the posterior is proportional to the product of the likelihood and the prior, to simplify our computation, we neglect the constant coefficients of related probability density functions in our derivation.

From Equation (2) and Equation (1), we have the likelihood function and the prior to be

$$f(\boldsymbol{q} \mid \boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{P\theta} - \boldsymbol{q})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{P\theta} - \boldsymbol{q})\right], \quad \pi(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\Pi})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\Pi})\right]$$

Leveraging the Bayes's formula, we have $f(\boldsymbol{\theta} \mid \boldsymbol{q}) \propto f(\boldsymbol{q} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. It follows that (neglecting terms that do not contain $\boldsymbol{\theta}$)

$$-2\log f(\boldsymbol{\theta} \mid \boldsymbol{q}) \propto (\boldsymbol{P\theta} - \boldsymbol{q})^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{P\theta} - \boldsymbol{q}) + (\boldsymbol{\theta} - \boldsymbol{\Pi})^\top \mathbf{C}^{-1}(\boldsymbol{\theta} - \boldsymbol{\Pi}) \xrightarrow[\text{drop terms without } \boldsymbol{\theta}]{\text{completing the squares}} \boldsymbol{\theta}^\top \left[\boldsymbol{P}^\top \boldsymbol{\Omega}^{-1}\mathbf{P} + \boldsymbol{C}^{-1}\right]\boldsymbol{\theta} - 2\left(\boldsymbol{q}^\top \boldsymbol{\Omega}^{-1}\mathbf{P} + \boldsymbol{\Pi}^\top \mathbf{C}^{-1}\right)\boldsymbol{\theta}$$

By Lemma 0.2, we obtain

$$\boldsymbol{\theta} \mid \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{\nu}, \boldsymbol{H}^{-1}\right), \quad \boldsymbol{\nu} = \left[\boldsymbol{P}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}\right]^{-1}\left[\boldsymbol{P}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{q} + \boldsymbol{C}^{-1}\boldsymbol{\Pi}\right] \text{ and } \boldsymbol{H}^{-1} = \left[\boldsymbol{P}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{P} + \boldsymbol{C}^{-1}\right]^{-1}$$

as desired.

APT Model:

Leveraging the powerful APT model (Roll & Ross, 1980; Ross, 1976), the parameter vector $\boldsymbol{\theta}$ could be generalized to represent the means of unobservable latent factors in the APT model, which assumes a linear functional form:

$$\boldsymbol{r} = \boldsymbol{Xf} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \text{ and } \boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right), \sigma_i^2 > 0, \forall i \tag{3}$$

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 $\boldsymbol{X}$ is a (non-random) $n \times k$ matrix that is known before time t. The variable $\boldsymbol{f}$ in Equation (3) denotes a k-dimensional latent random vector process, and 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

$$\mathbb{E}[\boldsymbol{f}] = \boldsymbol{\mu}_f \text{ and } \mathbb{V}[\boldsymbol{f}] = \boldsymbol{F}$$

In the Black-Litterman-Bayes model, we choose

$$\boldsymbol{\theta} = \boldsymbol{\mu}_f, \quad \boldsymbol{f} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{F})$$

Likelihood Function (Views):

For our understanding, let's assume we are considering two latent factors: value and momentum. Typically, a quantitative portfolio manager might have views on individual factors: (1) a view on the value premium, and (2) another view on the momentum premium. It would be atypical for portfolio managers to have views on a combination (e.g. linear) of factors. Hence to keep things simple but still useful, we take the views equation to be:

$$\boldsymbol{\theta}=\boldsymbol{q}+\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega}=\operatorname{diag}\left(\omega_1^2, \ldots, \omega_k^2\right)$$

then, $\boldsymbol{q} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Omega})$ and the corresponding likelihood function is therefore

$$f(\boldsymbol{q} \mid \boldsymbol{\theta}) \propto \exp \left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{q})^{\top} \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{q})\right]=\prod_{i=1}^k \exp \left[-\frac{1}{2 \omega_i^2}\left(\theta_i-q_i\right)^2\right]$$

Prior:

What prior for $\boldsymbol{\theta}$ should we choose? First, let's set the prior $\pi(\boldsymbol{\theta})$ to be

$$\pi(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\xi}, \boldsymbol{V})$$

where $\boldsymbol{\xi} \in \mathbb{R}^k$ 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 $\boldsymbol{V}$, and once a prior is chosen, we instantly obtain the prior predictive distribution $p(\boldsymbol{r})$ for $\boldsymbol{r}$ using the APT model and an associated prior (benchmark) portfolio with holdings $\boldsymbol{h}_B$ where $\boldsymbol{h}_B$ maximizes expected utility of wealth, where the expectation is taken with respect to the a priori distribution on asset returns

$$p(\boldsymbol{r})=\int p(\boldsymbol{r} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d \boldsymbol{\theta}, \quad \boldsymbol{h}_B=\arg \max _{\boldsymbol{h}} \mathbb{E}\left[u\left(\boldsymbol{h}^{\top} \boldsymbol{r}\right)\right]=\arg \max _{\boldsymbol{h}} \int u\left(\boldsymbol{h}^{\top} \boldsymbol{r}\right) p(\boldsymbol{r}) d \boldsymbol{r}$$

Let's see what our prior (benchmark) portfolio will look like.

Since $\boldsymbol{r}=\boldsymbol{X} \boldsymbol{f}+\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$ and $\boldsymbol{f} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{F})$, we can equivalently rewrite it as $\boldsymbol{f}=\boldsymbol{\theta}+\boldsymbol{w}$ where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{F})$, and $\boldsymbol{w}$ is independent of $\boldsymbol{\theta}$. Then, by properties of multivariate normal distribution, we have

$$\boldsymbol{r}=\boldsymbol{X} \boldsymbol{f}+\boldsymbol{\epsilon}=\boldsymbol{X} \boldsymbol{\theta}+\boldsymbol{X} \boldsymbol{w}+\boldsymbol{\epsilon} \quad \Longrightarrow \quad \boldsymbol{r}_\pi \sim \mathcal{N}\left(\boldsymbol{X} \boldsymbol{\xi}, \boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{X} \boldsymbol{F} \boldsymbol{X}^{\top}+\boldsymbol{D}\right)$$

The a priori optimal portfolio under CARA utility, $u(x)=-e^{-\lambda x}$, $\lambda \neq 0$, is then

$$\boldsymbol{h}_B:=\boldsymbol{h}^*=\left(\lambda \mathbb{V}_\pi[\boldsymbol{r}]\right)^{-1} \mathbb{E}_\pi[\boldsymbol{r}]=\lambda^{-1}\left(\boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{X} \boldsymbol{F} \boldsymbol{X}^{\top}+\boldsymbol{D}\right)^{-1} \boldsymbol{X} \boldsymbol{\xi}$$

Remark 1.2. In the original paper, the author derived the prior predictive mean and variance as

$$\mathbb{E}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{V}^{-1} \boldsymbol{\xi} \text{ and } \mathbb{V}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1}$$

where $\boldsymbol{H}:=\boldsymbol{V}^{-1}+\boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$ and $\boldsymbol{\Sigma}:=\boldsymbol{D}+\boldsymbol{X} \boldsymbol{F} \boldsymbol{X}^{\top}$.

We first note that due to his careless mistake, these results are wrong! The correct results are:

$$\mathbb{E}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{V}^{-1} \boldsymbol{\xi} \quad \text{and} \quad \mathbb{V}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1}$$

Then, substitute $\mathbf{A}=\boldsymbol{\Sigma}$, $\mathbf{U}=\mathbf{X}$, $\mathbf{V}=\mathbf{X}^{\top}$ and $\mathbf{B}=\mathbf{V}$ in the Woodbury identity (Theorem 0.1), we have

$$\mathbb{V}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1}=\boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{\Sigma}=\boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{X} \boldsymbol{F} \boldsymbol{X}^{\top}+\boldsymbol{D}$$

It follows that

$$\left(\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}=\left(\boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{\Sigma}\right) \boldsymbol{\Sigma}^{-1} \boldsymbol{X}=\boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}+\boldsymbol{X}=\boldsymbol{X} \boldsymbol{V}\left(\boldsymbol{V}^{-1}+\boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}\right)=\boldsymbol{X} \boldsymbol{V} \boldsymbol{H}$$

The mean $\mathbb{E}_\pi[\boldsymbol{r}]$ then follows as

$$\mathbb{E}_\pi[\boldsymbol{r}]=\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{X}^{\top} \boldsymbol{\Sigma}^{-1}\right)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \boldsymbol{H}^{-1} \boldsymbol{V}^{-1} \boldsymbol{\xi}=\boldsymbol{X} \boldsymbol{V} \boldsymbol{H} \boldsymbol{H}^{-1} \boldsymbol{V}^{-1} \boldsymbol{\xi}=\boldsymbol{X} \boldsymbol{\xi}$$

Thus, we obtain

$$\boldsymbol{r}_\pi \sim \mathcal{N}\left(\boldsymbol{X} \boldsymbol{\xi}, \boldsymbol{X} \boldsymbol{V} \boldsymbol{X}^{\top}+\boldsymbol{X} \boldsymbol{F} \boldsymbol{X}^{\top}+\boldsymbol{D}\right)$$

which is consistent with our results. △

(b) **Solution:** In this question, we derive the posterior distribution for $\boldsymbol{r} \mid \boldsymbol{q}$ and the mean-variance optimal portfolio holdings.

Step 1: Calculate the posterior for $\theta \mid q$.

Since $\pi(\boldsymbol{\theta})$ is normal and the likelihood $f(\boldsymbol{q} \mid \boldsymbol{\theta})$ is also normal, the prior is a conjugate prior with respect to the normal likelihood. Then, the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{q})$ is also normal. Using this relation and $p(\boldsymbol{\theta} \mid \boldsymbol{q}) \propto f(\boldsymbol{q} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})$, we have (neglecting terms that do not contain $\boldsymbol{\theta}$)

$$-2 \log p(\boldsymbol{\theta} \mid \boldsymbol{q})=(\boldsymbol{\theta}-\boldsymbol{q})^{\top} \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{q})+(\boldsymbol{\theta}-\boldsymbol{\xi})^{\top} \mathbf{V}^{-1}(\boldsymbol{\theta}-\boldsymbol{\xi})=\boldsymbol{\theta}^{\top}\left(\boldsymbol{\Omega}^{-1}+\mathbf{V}^{-1}\right) \boldsymbol{\theta}-2\left(\boldsymbol{q}^{\top} \boldsymbol{\Omega}^{-1}+\boldsymbol{\xi}^{\top} \mathbf{V}^{-1}\right) \boldsymbol{\theta}+(\text{ terms without } \boldsymbol{\theta})$$

Setting $\mathbf{H}=\left(\boldsymbol{\Omega}^{-1}+\mathbf{V}^{-1}\right)$ and $\boldsymbol{\eta}^{\top}=\left(\boldsymbol{q}^{\top} \boldsymbol{\Omega}^{-1}+\boldsymbol{\xi}^{\top} \mathbf{V}^{-1}\right)$ in Lemma 0.2, we have

$$\mathbb{V}[\boldsymbol{\theta} \mid \boldsymbol{q}]=\mathbf{H}^{-1}=\left(\boldsymbol{\Omega}^{-1}+\boldsymbol{V}^{-1}\right)^{-1}, \quad \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{q}]=\mathbf{H}^{-1} \boldsymbol{\eta}=\mathbb{V}[\boldsymbol{\theta} \mid \boldsymbol{q}]\left(\boldsymbol{V}^{-1} \boldsymbol{\xi}+\boldsymbol{\Omega}^{-1} q\right)=\left(\boldsymbol{V}^{-1}+\boldsymbol{\Omega}^{-1}\right)^{-1}\left(\boldsymbol{V}^{-1} \boldsymbol{\xi}+\boldsymbol{\Omega}^{-1} \boldsymbol{q}\right)$$

Therefore, the posterior must satisfy

$$\boldsymbol{\theta} \mid \boldsymbol{q} \sim \mathcal{N}(\widetilde{\boldsymbol{\xi}}, \widetilde{\boldsymbol{V}})$$

where $\widetilde{\boldsymbol{V}}=\left(\boldsymbol{V}^{-1}+\boldsymbol{\Omega}^{-1}\right)^{-1}$ and $\widetilde{\boldsymbol{\xi}}=\left(\boldsymbol{V}^{-1}+\boldsymbol{\Omega}^{-1}\right)^{-1}\left(\boldsymbol{V}^{-1} \boldsymbol{\xi}+\boldsymbol{\Omega}^{-1} \boldsymbol{q}\right)$.

Step 2: Calculate the posterior for $r \mid q$.

The a posteriori distribution of asset returns $\boldsymbol{r}$ (also called the posterior predictive density) is given by

$$p(\boldsymbol{r} \mid \boldsymbol{q}) = \int p(\boldsymbol{r} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{q}) d\boldsymbol{\theta}$$

For this one, we can view $p(\boldsymbol{\theta} \mid \boldsymbol{q})$ as the prior density for $\boldsymbol{\theta}$, then the situation here is the same as in deriving the prior (benchmark) portfolio. Hence, use the previous result and make the substitution $\boldsymbol{\xi} \to \widetilde{\boldsymbol{\xi}}$ and $\boldsymbol{V} \to \widetilde{\boldsymbol{V}}$, we obtain

$$\boldsymbol{r} \mid \boldsymbol{q} \sim \mathcal{N}\left(\boldsymbol{X}\widetilde{\boldsymbol{\xi}}, \boldsymbol{X}\widetilde{\boldsymbol{V}}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{D}\right)$$

Step 3: Calculate the mean-variance optimal portfolio.

The posterior optimal portfolio under CARA utility is then

$$\boldsymbol{h}^* = (\lambda \mathbb{V}[\boldsymbol{r} \mid \boldsymbol{q}])^{-1} \mathbb{E}[\boldsymbol{r} \mid \boldsymbol{q}] = \lambda^{-1} \left(\boldsymbol{X}\widetilde{\boldsymbol{V}}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{D}\right)^{-1} \boldsymbol{X}\widetilde{\boldsymbol{\xi}}$$

where $\widetilde{\boldsymbol{V}} = \left(\boldsymbol{V}^{-1} + \boldsymbol{\Omega}^{-1}\right)^{-1}$ and $\widetilde{\boldsymbol{\xi}} = \left(\boldsymbol{V}^{-1} + \boldsymbol{\Omega}^{-1}\right)^{-1} \left(\boldsymbol{V}^{-1}\boldsymbol{\xi} + \boldsymbol{\Omega}^{-1}\boldsymbol{q}\right)$.

2 Topic: Important Properties of the Kalman Filter

State Space Models:

Consider the following discrete linear dynamical system, which also referred to as linear state space model:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}_t \mathbf{x}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t \end{aligned} \quad \text{where} \quad \mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{P}_0), \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t), \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t) \tag{4}$$

Additionally, the initial state $\mathbf{x}_0$ and the noise terms $\mathbf{w}_t, \mathbf{v}_t$ are all assumed to be mutually independent.

Sequences $\{\mathbf{x}_\tau\}_{\tau=0}^t$ and $\{\mathbf{y}_\tau\}_{\tau=0}^t$ are called the hidden/latent state and observation sequences, respectively. It follows that the first and second equations are referred to the state and observation equations, respectively. In a state space model the state sequences $\{\mathbf{x}_\tau\}_{\tau=0}^t$ are not observable; however, the observation sequences $\{\mathbf{y}_\tau\}_{\tau=0}^t$ are fully observable.

The Filtering Problem:

The filtering problem is to estimate

$$\hat{\mathbf{x}}_t = \mathbb{E}(\mathbf{x}_t \mid \mathbf{Y}_t)$$

where $\mathbf{Y}_t := \{\mathbf{y}_0, \ldots, \mathbf{y}_t\}$. In other words, given the noisy observations, $\{\mathbf{y}_\tau\}_{\tau=0}^t$, we seek to estimate the expected state. And the optimal solution for our setup is given by the Kalman filter, also referred to as linear quadratic estimation (LQE).

Derive the Kalman Filter:

The derivation of the Kalman filter is by induction, showing that the observation and state sequences are conditionally Gaussian.

Base Case: Since $\mathbf{x}_0$ is Gaussian, $\mathbf{x}_1, \mathbf{y}_1$ are also Gaussians. It follows that

$$\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}\left(\mathbf{F}_1\boldsymbol{\mu}_0, \mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^\top + \mathbf{Q}_1\right) \\ \mathbf{y}_1 &\sim \mathcal{N}\left(\mathbf{H}_1\mathbf{F}_1\boldsymbol{\mu}_0, \mathbf{H}_1\left(\mathbf{F}_1\mathbf{P}_0\mathbf{F}_1^\top + \mathbf{Q}_1\right)\mathbf{H}_1^\top + \mathbf{R}_1\right) \end{aligned}$$

Induction Step: To perform the induction step, we assume that

$$\mathbf{x}_{t-1} \sim \mathcal{N}\left(\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right)$$

where the subscript $t \mid t'$ to denote our belief about the state $\mathbf{x}_t$ given observations $\mathbf{Y}_{t'}$. We immediately obtain from Equation (4)

$$\begin{aligned} \mathbf{x}_t \mid \mathbf{Y}_{t-1} &\sim \mathcal{N}\left(\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right) \\ \mathbf{y}_t \mid \mathbf{Y}_{t-1} &\sim \mathcal{N}\left(\mathbf{H}_t\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{H}_t\left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)\mathbf{H}_t^\top + \mathbf{R}_t\right) \end{aligned}$$

As $\mathbf{x}_t$, $\mathbf{w}_t$ and $\mathbf{v}_t$ are independent, the covariance of $\mathbf{x}_t$ and $\mathbf{y}_t$ is given by

$$\operatorname{Cov}(\mathbf{x}_t, \mathbf{y}_t) = \operatorname{Cov}(\mathbf{x}_t, \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t) = \operatorname{Cov}(\mathbf{x}_t, \mathbf{H}_t\mathbf{x}_t) + \operatorname{Cov}(\mathbf{x}_t, \mathbf{v}_t) = \mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^\top\mathbf{H}_t^\top\right] = \left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)\mathbf{H}_t^\top$$

IS/Predict Step of the Kalman Filter: From the distribution of $\mathbf{x}_t \mid \mathbf{Y}_{t-1}$, we introduce the notation

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &:= \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &:= \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t \end{aligned}$$

Then, we see that the joint distribution of $\mathbf{x}_t$ and $\mathbf{y}_t$, conditioned on $\mathbf{Y}_{t-1}$, is given by the multivariate Gaussian

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\mathbf{x}}_{t|t-1} \\ \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{H}_t^\top \\ \mathbf{H}_t\mathbf{P}_{t|t-1} & \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t \end{bmatrix}\right)$$

IS/Update Step of the Kalman Filter: Let us now assume $\mathbf{y}_t$ is observed. Applying the conditional formulas for the multivariate Gaussian we obtain

$$\mathbf{x}_t \mid \mathbf{Y}_t \sim \mathcal{N}\left(\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right)$$

where

$$\hat{\mathbf{x}}_{t|t} := \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1}\right), \quad \mathbf{P}_{t|t} := \left(\mathbf{I} - \mathbf{K}_t\mathbf{H}_t\right)\mathbf{P}_{t|t-1}, \quad \mathbf{K}_t := \mathbf{P}_{t|t-1}\mathbf{H}_t^\top\mathbf{S}_t^{-1}, \quad \mathbf{S}_t := \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t$$

Specifically, $\mathbf{K}_t$ is called the Kalman gain. Clearly, our belief about $\mathbf{x}_t$ given $\mathbf{Y}_t$ is a multivariate Gaussian distribution.

Remark 2.1. From the above filter formulas, it follows that we can apply the predict and update steps recursively as new observations arrive. This makes the Kalman filter ideal for online real-time processing of information. △

Problem 2.2. *Choose three out of the four subproblems. It is optional to solve the other one for extra credit.*

(a) *In class we derived the Kalman filter under the assumption that $\mathbb{E}\left[\mathbf{w}_t\mathbf{v}_t^\top\right] = 0$. Now, derive the Kalman filter under the assumption that $\mathbb{E}\left[\mathbf{w}_t\mathbf{v}_t^\top\right] = \mathbf{M}_t$.*

Solution: We assume that

$$\mathbf{x}_{t-1} \sim \mathcal{N}\left(\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right)$$

where the subscript $t \mid t'$ to denote our belief about the state $\mathbf{x}_t$ given observations $\mathbf{Y}_{t'}$.

Find $\mathbf{x}_t \mid \mathbf{Y}_{t-1}$ and $\mathbf{y}_t \mid \mathbf{Y}_{t-1}$:

From Equation (4), we get $\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{w}_t$. Then, for $\mathbf{x}_t \mid \mathbf{Y}_{t-1}$, it's easy to deduce that

$$\mathbf{x}_t \mid \mathbf{Y}_{t-1} \sim \mathcal{N}\left(\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)$$

Since $\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t = \mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t$, we have $\mathbb{E}\left[\mathbf{y}_t \mid \mathbf{Y}_{t-1}\right] = \mathbf{H}_t\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}$. By Lemma 0.4, we obtain

$$\begin{aligned}
\mathbb{V}\left[\mathbf{y}_t \mid \mathbf{Y}_{t-1}\right] = \mathbb{V}\left[\mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right] &= \operatorname{Cov}\left(\mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t, \mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right) \\
\mathbf{x}_{t-1} \text{ independent of } \mathbf{w}_t \text{ and } \mathbf{v}_t \Longrightarrow &= \operatorname{Cov}\left(\mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1}, \mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1}\right) + \operatorname{Cov}\left(\mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t, \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right) \\
&= \mathbf{H}_t\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top\mathbf{H}_t^\top + \mathbb{E}\left[\left(\mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right)\left(\mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right)^\top\right] \\
&= \mathbf{H}_t\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top\mathbf{H}_t^\top + \mathbf{H}_t\mathbb{E}\left[\mathbf{w}_t\mathbf{w}_t^\top\right]\mathbf{H}_t^\top + \mathbb{E}\left[\mathbf{v}_t\mathbf{v}_t^\top\right] + \mathbb{E}\left[\mathbf{H}_t\mathbf{w}_t\mathbf{v}_t^\top\right] + \mathbb{E}\left[\mathbf{v}_t\mathbf{w}_t^\top\mathbf{H}_t^\top\right] \\
&= \mathbf{H}_t\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top\mathbf{H}_t^\top + \mathbf{H}_t\mathbf{Q}_t\mathbf{H}_t^\top + \mathbf{R}_t + \mathbf{H}_t\mathbf{M}_t + \mathbf{M}_t^\top\mathbf{H}_t^\top \\
&= \mathbf{H}_t\left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)\mathbf{H}_t^\top + \mathbf{R}_t + \mathbf{H}_t\mathbf{M}_t + \mathbf{M}_t^\top\mathbf{H}_t^\top
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbf{x}_t \mid \mathbf{Y}_{t-1} &\sim \mathcal{N}\left(\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right) \\
\mathbf{y}_t \mid \mathbf{Y}_{t-1} &\sim \mathcal{N}\left(\mathbf{H}_t\mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{H}_t\left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)\mathbf{H}_t^\top + \mathbf{R}_t + \mathbf{H}_t\mathbf{M}_t + \mathbf{M}_t^\top\mathbf{H}_t^\top\right)
\end{aligned}$$

Find $\mathbf{x}_t \mid \mathbf{Y}_t$:

As $\mathbf{x}_{t-1}$, $\mathbf{w}_t$ and $\mathbf{x}_{t-1}$, $\mathbf{v}_t$ are independent, the covariance of $\mathbf{x}_t$ and $\mathbf{y}_t$ is given by

$$\operatorname{Cov}\left(\mathbf{x}_t, \mathbf{y}_t\right) = \operatorname{Cov}\left(\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{w}_t, \mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right) = \operatorname{Cov}\left(\mathbf{F}_t\mathbf{x}_{t-1}, \mathbf{H}_t\mathbf{F}_t\mathbf{x}_{t-1}\right) + \operatorname{Cov}\left(\mathbf{w}_t, \mathbf{H}_t\mathbf{w}_t + \mathbf{v}_t\right) = \left(\mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t\right)\mathbf{H}_t^\top + \mathbf{M}_t$$

- **IS/Predict Step of the Kalman Filter:** From the distribution of $\mathbf{x}_t \mid \mathbf{Y}_{t-1}$, we introduce the notation

$$\begin{aligned}
\hat{\mathbf{x}}_{t|t-1} &:= \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} \\
\mathbf{P}_{t|t-1} &:= \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^\top + \mathbf{Q}_t
\end{aligned}$$

Then, we see that the joint distribution of $\mathbf{x}_t$ and $\mathbf{y}_t$, conditioned on $\mathbf{Y}_{t-1}$, is given by the multivariate Gaussian

$$\begin{bmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\mathbf{x}}_{t|t-1} \\ \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{M}_t \\ \mathbf{H}_t\mathbf{P}_{t|t-1} + \mathbf{M}_t^\top & \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t + \mathbf{H}_t\mathbf{M}_t + \mathbf{M}_t^\top\mathbf{H}_t^\top \end{bmatrix}\right)$$

- **IS/Update Step of the Kalman Filter:** Let us now assume $\mathbf{y}_t$ is observed. Applying the conditional formulas for the multivariate Gaussian we obtain

$$\mathbf{x}_t \mid \mathbf{Y}_t \sim \mathcal{N}\left(\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right)$$

where

$$\hat{\mathbf{x}}_{t|t} := \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1}\right), \quad \mathbf{P}_{t|t} := \left(\mathbf{I} - \mathbf{K}_t\mathbf{H}_t\right)\mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{M}_t^\top, \quad \mathbf{K}_t := \mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{M}_t, \quad \mathbf{S}_t := \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t + \mathbf{H}_t\mathbf{M}_t + \mathbf{M}_t^\top\mathbf{H}_t^\top$$

Specifically, $\mathbf{K}_t$ is called the Kalman gain. Clearly, our belief about $\mathbf{x}_t$ given $\mathbf{Y}_t$ is a multivariate Gaussian distribution.

(b) *Using the notation from class, let us denote the error between the true and estimated states by the random variable $\widetilde{\mathbf{x}}_t = \mathbf{x}_t - \hat{\mathbf{x}}_t$. We denote by $\mathbf{W}_t$ a known positive definite matrix. Show that the Kalman filter is the solution to the problem*

$$\min_{\hat{\mathbf{x}}_t} \mathbb{E}\left[\widetilde{\mathbf{x}}_t^\top\mathbf{W}_t\widetilde{\mathbf{x}}_t\right] \tag{5}$$

Solution: Since $\mathbf{W}_t$ is a known positive definite matrix, by Cholesky decomposition, we have $\mathbf{W}_t = \mathbf{L}_t\mathbf{L}_t^\top$. Recall that the Kalman filter approximates $\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t]$ as $\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t|t}$, and using the fact that $\widetilde{\mathbf{x}}_t^\top \mathbf{W}_t \widetilde{\mathbf{x}}_t \in \mathbb{R}$, we can rewrite the loss function as

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\mathbf{x}}_t^\top \mathbf{W}_t \widetilde{\mathbf{x}}_t\right] &= \mathbb{E}\left[(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])^\top \mathbf{W}_t (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\right] + \mathbb{E}\left[(\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)^\top \mathbf{W}_t (\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)\right] \\
&\quad + 2\cdot\mathbb{E}\left[(\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)^\top \mathbf{W}_t (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\right] \\
&= \mathbb{E}\left[(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])^\top \mathbf{L}_t\mathbf{L}_t^\top (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\right] + \mathbb{E}\left[(\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)^\top \mathbf{L}_t\mathbf{L}_t^\top (\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)\right] \\
\phi(\hat{\mathbf{x}}_t) = (\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)^\top \mathbf{W}_t \implies &\quad + 2\mathbb{E}\left\{\mathbb{E}\left[\phi(\hat{\mathbf{x}}_t)(\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t]) \mid \mathbf{Y}_t\right]\right\} \\
&= \mathbb{E}\left[\|\mathbf{L}_t^\top (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\|^2\right] + \mathbb{E}\left[\|\mathbf{L}_t^\top (\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)\|^2\right] + 2\mathbb{E}\left\{\mathbb{E}\left[\phi(\hat{\mathbf{x}}_t)(\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\right]\right\} \\
&= \mathbb{E}\left[\|\mathbf{L}_t^\top (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\|^2\right] + \mathbb{E}\left[\|\mathbf{L}_t^\top (\mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t] - \hat{\mathbf{x}}_t)\|^2\right] \\
&\geq \mathbb{E}\left[\|\mathbf{L}_t^\top (\mathbf{x}_t - \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t])\|^2\right]
\end{aligned}$$

where the equality can be attained when $\hat{\mathbf{x}}_t = \mathbb{E}[\mathbf{x}_t \mid \mathbf{Y}_t]$.

Therefore, the Kalman filter is the solution to the problem $\min_{\hat{\mathbf{x}}_t} \mathbb{E}\left[\widetilde{\mathbf{x}}_t^\top \mathbf{W}_t \widetilde{\mathbf{x}}_t\right]$.

(c) ***(The Kalman filter is the Best LINEAR Predictor under Mean-Squared Error if noise is NOT Gaussian (Non-linear Predictors could be better.))*** *Now let us assume that $\mathbf{w}_t$ and $\mathbf{v}_t$ have zero mean, are uncorrelated with covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$, respectively (but they are no longer Gaussian). Show that the Kalman filter is the best linear solution to Equation (5).*

Solution: We assume that the estimate is a linear weighted sum of the prediction and the new observation and can be described by the equation,

$$\hat{\mathbf{x}}_{t|t} := \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1}\right)$$

where $\mathbf{K}_t$ is called the gain matrice. Our problem now is to reduced to finding the $\mathbf{K}_t$ that minimise the conditional mean squared estimation error where of course the estimation error is given by:

$$\widetilde{\mathbf{x}}_{t|t} = \mathbf{x}_t - \hat{\mathbf{x}}_{t|t}$$

Unbiasedness: Assume that at state $t-1$, our predictor $E\left[\hat{\mathbf{x}}_{t-1|t-1}\right] = E[\mathbf{x}_{t-1}]$ is unbiased. Notice that

$$E\left[\hat{\mathbf{x}}_{t|t-1}\right] = E\left[\mathbf{F}_{t-1}\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{w}_{t-1}\right] = \mathbf{F}_{t-1}E\left[\hat{\mathbf{x}}_{t-1|t-1}\right] = E[\mathbf{x}_t]$$

Taking the expectation yields

$$E\left[\hat{\mathbf{x}}_{t|t}\right] = E\left[\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1}\right)\right] = E[\mathbf{x}_t] + \mathbb{E}\left[\mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{K}_t\mathbf{v}_t - \mathbf{K}_t\mathbf{H}_t\hat{\mathbf{x}}_{t|t-1}\right] = E[\mathbf{x}_t]$$

Therefore, the result given by the Kalman filter formulated as above is unbiased provided $E\left[\hat{\mathbf{x}}_{t|t}\right] = E[\mathbf{x}_k]$.

Formulate the Error Covariance:

We now turn to the updated error covariance

$$E\left[\widetilde{\mathbf{x}}_{t|t}\widetilde{\mathbf{x}}_{t|t}^\top \mid \mathbf{Y}_k\right] = E\left[\left(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t}\right)\left(\mathbf{x}_k - \hat{\mathbf{x}}_{t|t}\right)^\top\right] = \mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\,\mathbf{P}_{t|t-1}$$

Our goal is now to minimise the conditional mean-squared estimation error with respect to the Kalman gain, $\mathbf{K}$.

$$L = \min_{\mathbf{K}_t} E\left[\widetilde{\mathbf{x}}_{t|t}^\top\widetilde{\mathbf{x}}_{t|t} \mid \mathbf{Y}_t\right] = \min_{\mathbf{K}_k}\operatorname{trace}\left(E\left[\widetilde{\mathbf{x}}_{t|t}\widetilde{\mathbf{x}}_{t|t}^\top \mid \mathbf{Y}_t\right]\right) = \min_{\mathbf{K}_k}\operatorname{trace}\left(\mathbf{P}_{t|t}\right)$$

For any matrix $\mathbf{A}$ and a symmetric matrix $\mathbf{B}$

$$\frac{\partial}{\partial\mathbf{A}}\left(\operatorname{trace}\left(\mathbf{A}\mathbf{B}\mathbf{A}^T\right)\right) = 2\mathbf{A}\mathbf{B}$$

(to see this, consider writing the trace as $\sum_i \mathbf{a}_i^T\mathbf{B}\mathbf{a}_i$ where $\mathbf{a}_i$ are the columns of $\mathbf{A}^T$, and then differentiating w.r.t. the $\mathbf{a}_i$).

Differentiating with respect to the gain matrix (using the relation above) and setting equal to zero yields

$$\frac{\partial L}{\partial\mathbf{K}_t} = -2(\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\,\mathbf{P}_{t|t-1}\mathbf{H}_t^T + 2\mathbf{K}_t\mathbf{R}_t = \mathbf{0}$$

Re-arranging gives an equation for the gain matrix

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^T\left[\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^T + \mathbf{R}_t\right]^{-1}$$

This is exactly what we have. Thus, the result given by the Kalman filter formulated as above is the best unbiased linear predictor provided $E\left[\hat{\mathbf{x}}_{t|t}\right] = E[\mathbf{x}_k]$.

(d) *In class we derived the Kalman filter under the assumption that $\mathbf{w}_t$ and $\mathbf{v}_t$ are uncolored (i.e. each is serially uncorrelated).*

(i) *Derive the Kalman filter under the assumption that $\mathbf{w}_t$ is a VAR(1) process with known system matrix.*

(ii) *Derive the Kalman filter under the assumption that $\mathbf{v}_t$ is a $\mathrm{VAR}(1)$ process with known system matrix.*

Hint: For (d), both (i) and (ii) can be solved by properly augmenting the state equations. Can you find a solution to (ii) where one does not have to augment the state? Is it possible to do so for (i) - why, or why not?

Solution:

3 Topic: Kalman filter's Application on Financial Problems — Pairs Trading

Problem 3.1. *In this question we consider a basic pairs trading strategy between two stocks with prices p_t^A and p_t^B at time t. We denote the spread between them by $s_t := \log\left(p_t^A\right) - \log\left(p_t^B\right)$ and assume the spread follows an Ornstein-Uhlenbeck process*

$$ds_t = \kappa\left(\theta - s_t\right)dt + \sigma dB_t \tag{6}$$

where dB_t is a standard Brownian motion. In other words, the spread reverts to its mean $\theta \in \mathbb{R}$ at the speed $\kappa \in \mathbb{R}_+$and volatility $\sigma \in \mathbb{R}_+$.

(a) *In the discrete sense, we use the notation $S_n := S_{n\cdot\Delta t}$ to denote the value of process s at a given time $n \cdot \Delta t$ where Δt is a constant period amount of time picked as the step size. Show that the discrete time solution of Equation (6) is Markovian, that is*

$$s_n = \mathbb{E}\left[s_n \mid s_{n-1}\right] + \varepsilon_n$$

where $n = 1, 2, \ldots$, and ε_n is a random process with zero mean and variance equal to $\sigma^2_{\varepsilon,n} = \mathbb{V}\left[s_n \mid s_{n-1}\right]$.

Hint: You can derive the discrete solution explicitly.

Solution: Set $f(t,x) = e^{\kappa t}x$ and apply the Itô's lemma, we have $de^{e^{\kappa t}s_t} = \kappa e^{\kappa t}s_t dt + e^{\kappa t}ds_t$. Pluging in Equation (6), we obtain

$$ds_t + \kappa s_t dt = \frac{de^{\kappa t}s_t}{e^{\kappa t}} = \kappa\theta dt + \sigma dB_t \Longrightarrow e^{\kappa t}s_t - s_0 = \int_0^t e^{\kappa s}\kappa\theta ds + \int_0^t e^{\kappa s}\sigma dB_s \Longrightarrow s_t = e^{-\kappa t}s_0 + \theta\left(1 - e^{-\kappa t}\right) + \int_0^t e^{\kappa(s-t)}\sigma dB_s, \forall t > 0$$

To extract s_{n-1} out of s_n, we can rewrite s_n as:

$$\begin{aligned}
s_n &= e^{-\kappa n\Delta t}s_0 + \theta\left(1 - e^{-\kappa n\Delta t}\right) + \int_0^n e^{\kappa(s-n)\Delta t}\sigma dB_s \\
&= \theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}\cdot\left(e^{-\kappa(n-1)\Delta t}s_0 + \theta\left(1 - e^{-\kappa(n-1)\Delta t}\right) + e^{-\kappa(n-1)\Delta t}\left[\int_0^{(n-1)\Delta t} e^{\kappa s}\sigma dB_s + \int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s\right]\right) \\
&= \theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}s_{n-1} + e^{-\kappa n\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s
\end{aligned}$$

Taking the mean and variance of s_n conditioned on s_{n-1}, and define $\varepsilon_n := e^{-\kappa n\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s$, we have

$$\begin{aligned}
\mathbb{E}\left[s_n \mid s_{n-1}\right] &= \mathbb{E}\left[\theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}s_{n-1}\right] + \mathbb{E}\left[e^{-\kappa n\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s \mid s_{n-1}\right] \xlongequal[\text{information up to } n-1]{s_{n-1} \text{ only contains}} \theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}s_{n-1} \\
\mathbb{V}\left[s_n \mid s_{n-1}\right] &= \mathbb{V}\left[\theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}s_{n-1} + e^{-\kappa n\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s \mid s_{n-1}\right] = \mathbb{V}\left[e^{-\kappa n\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} e^{\kappa s}\sigma dB_s\right] = \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa\Delta t}\right) = \mathbb{V}\left[\varepsilon_n\right]
\end{aligned}$$

Therefore, we have

$$s_n = \mathbb{E}\left[s_n \mid s_{n-1}\right] + \varepsilon_n$$

where ε_n is a random process with zero mean and variance equal to $\sigma^2_{\varepsilon,n} = \mathbb{V}\left[s_n \mid s_{n-1}\right] = \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa\Delta t}\right)$.

(b) *Propose a methodology for updating the parameters θ, κ using the Kalman filter and describe how you would use it to trade the stock pair.*

Solution: To conform the notation used in the Kalman filter, we use $t \in \mathbb{N}$ to denote the states and write the result in (a) as:

$$s_t = \theta - e^{-\kappa\Delta t}\theta + e^{-\kappa\Delta t}s_{t-1} + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa\Delta t}\right)\right)$$

State Equation:

Set constants $\alpha = \theta - e^{-\kappa\Delta t}\theta$, $\beta = e^{-\kappa\Delta t}$, $\mathbf{F}_t = \begin{bmatrix}1 & 0\\ \alpha & \beta\end{bmatrix}$, $\mathbf{x}_t = \begin{bmatrix}1\\ s_t\end{bmatrix}$, $\mathbf{w}_t = \begin{bmatrix}0\\ \varepsilon_t\end{bmatrix}$ and $\mathbf{Q}_t = \begin{bmatrix}0 & 0\\ 0 & \sigma^2_{\epsilon,t}\end{bmatrix}$, we have our state equation as:

$$s_t = \alpha + \beta s_{t-1} + \varepsilon_t \xrightarrow{\text{rewrite}} \begin{bmatrix}1\\ s_t\end{bmatrix} = \begin{bmatrix}1 & 0\\ \alpha & \beta\end{bmatrix}\begin{bmatrix}1\\ s_{t-1}\end{bmatrix} + \begin{bmatrix}0\\ \varepsilon_t\end{bmatrix} \xrightarrow{\text{substitute } \mathbf{F}_t, \mathbf{x}_t \text{ and } \mathbf{w}_t} \mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{w}_t \quad \text{where} \quad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_t\right)$$

Observations Equation:

Now, assume the observation y_t is just the true process s_t plus a small perturbation. Setting $\mathbf{H}_t = \begin{bmatrix}0 & 1\end{bmatrix}$, $\mathbf{v}_t = v \sim \mathcal{N}\left(0, \nu^2\right)$ and $\mathbf{R}_t = \nu^2$, we obtain

$$y_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t \quad \text{where} \quad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}_t\right)$$

Initial State:

Naturally, we set the initial state variable $\mathbf{x}_0 \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{P}_0\right)$ where $\boldsymbol{\mu}_0 = \begin{bmatrix}1\\ s_0\end{bmatrix}$ and $\mathbf{P}_0 = \mathbf{0}$.

Therefore, by our construction, we have the linear state space model:

$$\begin{aligned}
\mathbf{x}_t &= \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{w}_t \\
\mathbf{y}_t &= \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t
\end{aligned} \quad \text{where} \quad \mathbf{x}_0 \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{P}_0\right), \quad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}_t\right), \quad \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}_t\right)$$

Additionally, the initial state $\mathbf{x}_0$ and the noise terms $\mathbf{w}_t, \mathbf{v}_t$ are all assumed to be mutually independent.

(c) *Test your methodology from (b) on simulated data. In particular, (i) simulate (2) from known parameters θ, κ and σ, and then (ii) use the Kalman filter to recover them. You do not need to implement the Kalman filter from scratch; you are welcome to use a Kalman implementation from a Python package such as pykalman. How do you obtain a good estimate of σ ?*

Solution:

(d) *Repeat the same experiment from (c), but this time simulate (2) first with κ having the same value as above and then suddenly changing it to another value such that the half-life of the spread is 50% of its original value. How long does it take the Kalman filter to adjust? Can you make adjustment to your filter in order to speed up the time it takes the Kalman filter to adjust?*

Solution:

Hint: For (c) and (d), think about how you are going to demonstrate the results using appropriate graphs, etc.

4 Topic: Kalman filter's Application on Financial Problems — Index Tracking Portfolios

Problem 4.1. *It is common in portfolio management to build so-called (index) tracking portfolios. Let us assume we are observing the return of the S&P 500 benchmark index, $r_{b,t}$. Now, let us pick a subset of 50 stocks from the constituents of this index. We will use these stocks to build a tracking portfolio for the index. For example, this could be the 50 companies in the index with the largest market cap. We denote the returns of these 50 stock by $\mathbf{r}_t \in \mathbb{R}^{50}$. The goal of finding a tracking portfolio is to find a dynamic trading strategy of the 50 stocks such that $\beta_t^\top \mathbf{r}_t \approx r_{b,t}$, where β_t denotes the holdings (relative weights) of the tracking portfolio.*

(a) *In this part, we assume that the covariance matrix of returns of the stocks in the S&P500, $\mathbf{\Sigma}$, is given and constant through time. Find the portfolio of these 50 stocks that minimizes the tracking error to $r_{b,t}$, i.e. find the solution to*

$$\boldsymbol{\beta}_t^* = \text{argmin}_{\boldsymbol{\beta}_t} \sqrt{\mathbb{V}\left[r_{b,t} - \boldsymbol{\beta}_t^\top \mathbf{r}_t\right]}.$$

What specific property does β_t^* have here?

Solution:

Formulation 1:

Suppose $\boldsymbol{S}$ and $\boldsymbol{\mu}_t = \mathbb{E}[\boldsymbol{r}_t]$ are the covariance matrix and the mean of the returns of the 50 stocks we picked, respectively. Additionally, let $\boldsymbol{e} = (1, \ldots, 1) \in \mathbb{R}^{50}$ denote the vector of all 1's. Since $r_{b,t}$ is deterministic, $\mathbb{V}\left[r_{b,t} - \boldsymbol{\beta}_t^\top \mathbf{r}_t\right] = \mathbb{V}\left[\boldsymbol{\beta}_t^\top \mathbf{r}_t\right] = \mathbb{E}\left[\boldsymbol{\beta}_t^\top \mathbf{r}_t \mathbf{r}_t^\top \boldsymbol{\beta}_t\right] = \boldsymbol{\beta}_t^\top \boldsymbol{S} \boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_t^\top \mathbf{r}_t \approx r_{b,t}$, we have

$$\boldsymbol{\beta}_t^* = \text{argmin}_{\boldsymbol{\beta}_t} \sqrt{\boldsymbol{\beta}_t^\top \boldsymbol{S} \boldsymbol{\beta}_t} = \text{argmin}_{\boldsymbol{\beta}_t} \frac{1}{2} \boldsymbol{\beta}_t^\top \boldsymbol{S} \boldsymbol{\beta}_t \quad \text{given} \quad \boldsymbol{\beta}_t^\top \mathbb{E}[\mathbf{r}_t] = \boldsymbol{\beta}_t^\top \boldsymbol{\mu}_t = r_{b,t}, \quad \boldsymbol{\beta}_t^\top \boldsymbol{e} = 1$$

Applying the Lagrange multipliers, we have Lagrangian

$$L \equiv L(\boldsymbol{\beta}_t, \lambda, \gamma) := \frac{1}{2} \boldsymbol{\beta}_t^\top \boldsymbol{S} \boldsymbol{\beta}_t + \lambda \left(1 - \boldsymbol{\beta}_t^\top \boldsymbol{e}\right) + \gamma \left(r_{b,t} - \boldsymbol{\beta}_t^\top \boldsymbol{\mu}_t\right)$$

The first-order necessary condition then yields

$$FOC: \frac{\partial L}{\partial \boldsymbol{\beta}_t} = \boldsymbol{S}\boldsymbol{\beta}_t - \lambda \boldsymbol{e} - \gamma \boldsymbol{\mu}_t = \mathbf{0} \quad \Longrightarrow \quad \boldsymbol{\beta}_t = \boldsymbol{S}^{-1}(\lambda \boldsymbol{e} + \gamma \boldsymbol{\mu}_t) \tag{7}$$

Since $\boldsymbol{\beta}_t^\top \boldsymbol{\mu}_t = r_{b,t}$ and $\boldsymbol{\beta}_t^\top \boldsymbol{e} = 1$, we obtain

$$\begin{aligned} e^\top \boldsymbol{S}^{-1}(\lambda e + \gamma \boldsymbol{\mu}_t) &= 1, \\ \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1}(\lambda e + \gamma \boldsymbol{\mu}_t) &= r_{b,t} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \boldsymbol{e}^\top \boldsymbol{S}^{-1} e \lambda + \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t \gamma &= 1, \\ \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} e \lambda + \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t \gamma &= r_{b,t} \end{aligned}$$

Now, for simplicity, we set $A = \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{e}, B = \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t = \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} \boldsymbol{e}$, and $C = \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t$ and obtain

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda \\ \gamma \end{bmatrix} = \begin{bmatrix} 1 \\ r_{b,t} \end{bmatrix} \Longrightarrow \begin{bmatrix} \lambda^* \\ \gamma^* \end{bmatrix} = \begin{bmatrix} A & B \\ B & C \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ r_{b,t} \end{bmatrix} = \frac{1}{AC - B^2} \begin{bmatrix} C & -B \\ -B & A \end{bmatrix} \begin{bmatrix} 1 \\ r_{b,t} \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} C - Br_{b,t} \\ -B + Ar_{b,t} \end{bmatrix} \tag{8}$$

where $\Delta = AC - B^2$. Then, plug in Equation (7), we get

$$\boldsymbol{\beta}_t^* = \lambda^* \boldsymbol{S}^{-1} \boldsymbol{e} + \gamma^* \boldsymbol{S}^{-1} \boldsymbol{\mu}_t$$

This is clearly the minimum as $\frac{\partial^2 L}{\partial \beta_t^2} = \boldsymbol{S}$ is positive definite.

We call $\boldsymbol{\beta}_t^*$ a mean-variance efficient portfolio/allocation with respect to $r_{b,t}$.

Properties: By construction, the expected return of the optimal portfolio is

$$r_{b,t} = {\boldsymbol{\beta}_t^*}^\top \boldsymbol{\mu}_t$$

Some algebra shows that the variance of the optimal portfolio is

$$\begin{aligned} \sigma_{b,t}^{*2} &= {\boldsymbol{\beta}_t^*}^\top \boldsymbol{S} \boldsymbol{\beta}_t^* \\ &= \left(\lambda^* \boldsymbol{S}^{-1} \boldsymbol{e} + \gamma^* \boldsymbol{S}^{-1} \boldsymbol{\mu}_t\right)^\top \boldsymbol{\Sigma} \left(\lambda^* \boldsymbol{S}^{-1} \boldsymbol{e} + \gamma^* \boldsymbol{S}^{-1} \boldsymbol{\mu}_t\right) \\ &= {\lambda^*}^2 \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{e} + 2\lambda^* \gamma^* \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t + {\gamma^*}^2 \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t \\ &= \lambda^* \left(\lambda^* \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{e} + \gamma^* \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t\right) + \gamma^* \left(\lambda^* \boldsymbol{e}^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t + \gamma^* \boldsymbol{\mu}_t^\top \boldsymbol{S}^{-1} \boldsymbol{\mu}_t\right) \\ &= \lambda^* \left(\lambda^* A + \gamma^* B\right) + \gamma^* \left(\lambda^* B + \gamma^* C\right) \\ \text{Equation (8)} \Longrightarrow &= \lambda^* \cdot 1 + \gamma^* \cdot r_{b,t} \\ \text{Equation (8)} \Longrightarrow &= \frac{1}{\Delta} \left(C - Br_{b,t}\right) + \frac{1}{\Delta} \left(-B + Ar_{b,t}\right) r_{b,t} \\ &= \frac{Ar_{b,t}^2 - 2Br_{b,t} + C}{\Delta} \end{aligned}$$

Hence, we have $\sigma_{b,t}^{*^2}$ as a quadratic function of $r_{b,t}$:

$$\sigma_{b,t}^{*^2} = \frac{Ar_{b,t}^2 - 2Br_{b,t} + C}{\Delta}$$

If we let $r_{b,t}$ vary, the efficient portfolios β_t^* form a a hyperbola in the $(\sigma^*(r_{b,t}), r_{b,t})$ plane and a parabola in the $\left(\sigma^{*^2}(r_{b,t}), r_{b,t}\right)$-plane, and we call it the efficient frontier.

Formulation 2:

Since minimizing the given loss function amounts to minimizing the MSE which falls under the context of linear regression, we can view this index tracking problem as a regression problem. Since the covaraince matrix is assumed to be constant over time, we can specify a rolling window, say 30 days, and treat pairs $\{(r_{b,t-i}, \mathbf{r}_{t-i})\}$, $i = 0, 1, \ldots, 29$ as samples. Let's denote $\mathbf{y}_t = (r_{b,t}, r_{b,t-1}, \ldots, r_{b,t-29}) \in \mathbb{R}^{30}$ as the regressand, and $\mathbf{R}_t = \begin{bmatrix} \mathbf{r}_t^\top \\ \vdots \\ \mathbf{r}_{t-29}^\top \end{bmatrix}$ as regressors. Then, for time t, we can assume

$$\mathbf{y}_t = \mathbf{R}_t \boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t)$$

By classical linear regression, we have

$$\boldsymbol{\beta}_t^* = \left(\mathbf{R}_t^\top \mathbf{R}_t\right)^{-1} \mathbf{R}_t^\top \mathbf{y}_t$$

By the Gauss-Markov theorem, $\boldsymbol{\beta}_t^*$ is the best unbiased linear predictor under the assumptions of classical linear regression.

(b) *In this part, we no longer assume that covariances amongst stocks are time invariant. Propose a solution to minimizing the tracking error using the Kalman filter.*

Solution: Leveraging the powerful Kalman filter, we can update our view on $\boldsymbol{\beta}_t$ from every single new observations on $r_{b,t}$.

State Equation & Observations Equation:

Set $\mathbf{x}_t = \boldsymbol{\beta}_t \in \mathbb{R}^{50}$, $\mathbf{F}_t = \mathbf{I}_{50} \in \mathbb{R}^{50\times 50}$, $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$, $\mathbf{Q}_t = \mathbf{0}$, $\mathbf{y}_t = r_{b,t} \in \mathbb{R}$, $\mathbf{H}_t = \mathbf{r}_t^\top \in \mathbb{R}^{1\times 50}$, $\mathbf{v}_t = v \sim \mathcal{N}(0, \mathbf{R}_t)$ and $\mathbf{R}_t = 0.3$.

Initial State:

Naturally, we set the initial state variable $\mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{P}_0)$ where $\boldsymbol{\mu}_0 = (1, \ldots, 1) \in \mathbb{R}^{50}$ and $\mathbf{P}_0 = \mathbf{I}_{50} * 0.1$.

Therefore, by our construction, we have the linear state space model:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}_t \mathbf{x}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t \end{aligned} \quad \text{where} \quad \mathbf{x}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{P}_0), \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t), \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$$

Additionally, the initial state $\mathbf{x}_0$ and the noise terms $\mathbf{w}_t, \mathbf{v}_t$ are all assumed to be mutually independent.

(c) *Download daily market data and create an example that illustrates your methodology. How does the tracking portfolio of the Kalman filter perform?*

Solution: By comparing errors to $r_{b,t}$, the Kalman filter performs a lot better than the rolling window regression scheme.

For detailed codes, please check the attached Jupyter notebook.

Hint: For (c), compare your Kalman filter to a solution based on (a) where the covariance matrix is estimated on rolling windows. Can you match the performance of the Kalman filter with the simpler methodology in (a) by appropriately choosing the length of the rolling window?

5 Topic: Another Interpretation for Independent Component Analysis (ICA)

Problem 5.1. *In class we considered the principally-independent component analysis method which essentially was the truncated rank- K SVD of a matrix X followed by an ICA rotation of the left singular components:*

$$\mathbf{X} \simeq \mathbf{U}\mathbf{S}\mathbf{V}^\top = \mathbf{U}_I \mathbf{S}_I \mathbf{V}_I$$

where $\mathbf{U}_I = \mathbf{U}\mathbf{A}_I$ *with* $\mathbf{A}_I = \underset{\mathbf{A}, \mathbf{A}^\top \mathbf{A} = I_K}{\operatorname{argmax}} |k_\ell(\mathbf{U}\mathbf{A})|, k_\ell(\mathbf{G})$ *being any centered cumulant of order* $\ell \geq 3$ *which for all practical purposes can be considered a non-linear (activation)*

function applied to each of the entries of $\mathbf{G}$. *Furthermore the matrix* $\mathbf{V}_I$ *was defined as* $\mathbf{V}_I^\top := \mathbf{D}^{-1}\mathbf{S}^{-1}\mathbf{A}_I\mathbf{S}\mathbf{V}^\top$ *where* $\mathbf{D}$ *was chosen so that* $\mathbf{V}_I$ *has unimodular columns.*

(a) *Show that* $\mathbf{D}$ *is a diagonal matrix*

(b) *Show that* $\mathbf{S}_I = \mathbf{S}\mathbf{D}$ *is diagonal such that* $\operatorname{Tr}(\mathbf{S}_I^2) = \operatorname{Tr}(\mathbf{S}^2)$.

(c) *Show that the method can be derived as the limit,* $\lambda^2 \to 0$, *of the optimization*

$$\mathbf{U}_I, \mathbf{S}_I, \mathbf{V}_I = \underset{\mathbf{P}, \mathbf{Q}: \mathbf{P}^\top \mathbf{P} = \operatorname{diag}(\mathbf{Q}^\top \mathbf{Q}) = \mathbf{I}_k}{\operatorname{argmin}} \left\| \mathbf{X} - \mathbf{P}\mathbf{R}\mathbf{Q}^\top \right\|_F - \lambda^2 |k_\ell(\mathbf{P})|.$$

(d) *Show that an alternative objective function achieving the same result is*

$$\mathbf{U}_I, \mathbf{S}_I, \mathbf{V}_I = \underset{\mathbf{P}, \mathbf{Q}: \mathbf{P}^\top \mathbf{P} = \operatorname{diag}(\mathbf{Q}^\top \mathbf{Q}) = \mathbf{I}_k}{\operatorname{argmin}} \left\| \mathbf{X} - \mathbf{P}\mathbf{R}\mathbf{Q}^\top \right\|_F - \lambda^2 J(\mathbf{P}),$$

where $J[\mathbf{x}] := H[\mathbf{x}_{gauss}] - H[\mathbf{x}]$ *is the negentropy and* $J(\mathbf{P})$ *is the sum of the negentropies of all the columns of* $\mathbf{P}$.

References

[1] Kolm, Petter N. and Ritter, Gordon, On the Bayesian Interpretation of Black-Litterman (October 16, 2016). European Journal of Operational Research, Volume 258, Issue 2, 16 April 2017, Pages 564-572, Available at SSRN: https://ssrn.com/abstract=2853158.