Optimization

5. Convex optimization: uses

Andrew Lesniewski

Baruch College
New York

Fall 2018

Outline

1. Parameter estimation

2. Portfolio optimization problems

Maximum likelihood estimation

- *Maximum likelihood estimation* (MLE) is a methodology used to estimate the parameters in a probability distribution from a set of observations.
- The observations $y_1, \ldots, y_N$ may come (but do not have to) in the form of a time series of asset prices, credit spreads, economic numbers, etc.
- We assume that the observations are independent draws from a parametrically given probability distribution $p(y|\theta)$, where $\theta_1, \ldots, \theta_n$ are the unknown parameters of the distribution.
- The underlying principle of MLE is that θ is chosen so that it maximizes the probability $p(y_1|\theta) \ldots p(y_N|\theta)$ of the observed set. In other words, we seek to *maximize* the *likelihood function*

$$\mathcal{L}(\theta|y) = \prod_{j=1}^{N} p(y_j|\theta). \tag{1}$$

- Equivalently, we seek to *minimize* the *log likelihood function* $-\log \mathcal{L}(\theta|y)$.

Maximum likelihood estimation

- MLE problems are, in general, not convex. In a number of lucky situations, they are.
- Consider the linear model discussed, whose special case was discussed in Lecture Notes #1 (using a slightly different notation). We assume that

$$y = Ax + \varepsilon, \tag{2}$$

where

(i) $x \in \mathbb{R}^n$ is the vector of unknown parameters (to be estimated),
(ii) $y \in \mathbb{R}^m$ is the vector of observations,
(iii) $\varepsilon \in \mathbb{R}^m$ is the vector of residuals; they are assumed to be random i.i.d. variables with a known parametric probability density function (PDF) $p(\varepsilon)$.

- The likelihood function of x given the observations y reads

$$\mathcal{L}(x|y) = \prod_{i=1}^{m} p\big(y_i - (Ax)_i\big), \tag{3}$$

and so the goal is to minimize the log likelihood function

$$-\log \mathcal{L}(x|y) = -\sum_{i=1}^{m} \log p\big(y_i - (Ax)_i\big). \tag{4}$$

Maximum likelihood estimation

- If the noise is Gaussian $\varepsilon \sim N(0, \sigma^2)$, i.e. $p(\varepsilon) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/2\sigma^2)$, then

  $$-\log \mathcal{L}(x|y) = \frac{m}{2} \log(2\pi^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{m} \big(y_i - (Ax)_i\big)^2, \tag{5}$$

  and the problem reduces to minimizing the L^2-norm

  $$\|y - Ax\|_2^2 = \sum_{i=1}^{m} \big(y_i - (Ax)_i\big)^2. \tag{6}$$

- If the noise is Laplace distributed, $p(\varepsilon) = (2a)^{-1} \exp(-|\varepsilon|/a)$, $a > 0$, then

  $$-\log \mathcal{L}(x|y) = m \log(2a) + \frac{1}{a} \sum_{i=1}^{m} \big|y_i - (Ax)_i\big|, \tag{7}$$

  and the problem reduces to minimizing the L^1-norm

  $$\|y - Ax\|_1 = \sum_{i=1}^{m} \big|y_i - (Ax)_i\big|. \tag{8}$$

Maximum likelihood estimation

- *Counting processes* are used to model event risk, such as credit default, loan prepayment, or trade fill.
- An example of a distribution used is the Poisson distribution, according to which the number of events Y has the distribution

$$\operatorname{Prob}(Y = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \tag{9}$$

  where $\lambda > 0$ is the event intensity.
- We may model the intensity as a linear function of a number of observable factors $x_1, \ldots, x_n$,

$$\mu = a^{\mathrm{T}} x + b \tag{10}$$

  and the goal is to estimate the parameters a and b.

Maximum likelihood estimation

- The likelihood function reads

$$\mathcal{L}(a, b|x, y) = \prod_{i=1}^{m} \frac{(a^{\mathrm{T}} x_i + b)^{y_i}}{y_i!}\, e^{-(a^{\mathrm{T}} x_i + b)}, \tag{11}$$

and thus the problem is to minimize the log likelihood function

$$-\log \mathcal{L}(a, b|x, y) = \sum_{i=1}^{m} \big(a^{\mathrm{T}} x_i + b - y_i \log(a^{\mathrm{T}} x_i + b) + \log y_i!\big). \tag{12}$$

- The optimal parameters a^* and b^* are thus the (unique) solutions to the following convex optimization problem:

$$\min \sum_{i=1}^{m} \big(a^{\mathrm{T}} x_i + b - y_i \log(a^{\mathrm{T}} x_i + b)\big). \tag{13}$$

Maximum likelihood estimation

- Oftentimes, in event modeling, the event has a binary outcome $Y \in \{0, 1\}$, and

$$\begin{aligned} \mathrm{Prob}(Y = 1) &= p, \\ \mathrm{Prob}(Y = 0) &= 1 - p. \end{aligned} \tag{14}$$

- The outcomes of such events can be modeled by means of *logistic regression*:

$$p = \frac{\exp(a^{\mathrm{T}} x + b)}{1 + \exp(a^{\mathrm{T}} x + b)} \tag{15}$$

where x is the observed vector of factors impacting the outcomes.

Maximum likelihood estimation

- Given a set of observations (x_i, y_i) (remember, $y_i \in \{0, 1\}$), we can write the likelihood functions as

$$\prod_{i:\, y_i=1} p_i \prod_{i:\, y_i=0} (1-p_i) = \prod_{i:\, y_i=1} \frac{\exp(a^{\mathrm{T}} x_i + b)}{1+\exp(a^{\mathrm{T}} x_i + b)} \prod_{i:\, y_i=0} \frac{1}{1+\exp(a^{\mathrm{T}} x_i + b)} \,. \tag{16}$$

- Minimization of the log likelihood function

$$-\log \mathcal{L}(a, b|x, y) = \sum_{i=1}^{m} \log\left(1+\exp(a^{\mathrm{T}} x_i + b)\right) - \sum_{i:\, y_i=1} (a^{\mathrm{T}} x_i + b) \tag{17}$$

is convex, and so the problem can be solved by means of convex optimization.

Maximum likelihood estimation

- Suppose that Y is an n-dimensional Gaussian random variable, and our task is to estimate the covariance matrix C of Y from a set of observations $y_i \in \mathbb{R}^n$, $i = 1, \ldots, N$.
- The PDF of Y is $p(y) = (2\pi)^{-n/2} \det(C)^{-1/2} \exp\big(-\frac{1}{2}\, y^{\mathrm{T}} C^{-1} y \big)$, and so the log likelihood function is given by

$$\begin{aligned} -\log \mathcal{L}(C|y_1, \ldots, y_N) &= \frac{1}{2} \sum_{j=1}^{N} y_i^{\mathrm{T}} C^{-1} y_i + \frac{N}{2} \log\det(C) + \frac{Nn}{2} \log(2\pi) \\ &= \frac{N}{2} \operatorname{tr}\big(\widehat{C} C^{-1}\big) + \frac{N}{2} \log\det(C) + \frac{Nn}{2} \log(2\pi), \end{aligned}$$

where

$$\widehat{C} = \frac{1}{N} \sum_{j=1}^{N} y_i y_i^{\mathrm{T}}$$

is the sample estimate of the covariance.

Maximum likelihood estimation

- The optimization problem is thus:

$$\min \ \operatorname{tr}\big(\widehat{C}C^{-1}\big) + \log\det(C) \quad \text{subject to } C \in \mathbb{P}^n_+, \ C \text{ invertible.}$$

- It is sometimes convenient to substitute $S = C^{-1}$ (S is called the *precision matrix*, and formulate the problem as

$$\min \ \operatorname{tr}\big(\widehat{C}S\big) - \log\det(S) \quad \text{subject to } S \in \mathbb{P}^n_n, \ S \text{ invertible.}$$

- The problem has an explicit solution $C^* = \widehat{C}$, i.e. the sample covariance is the MLE estimate of the covariance matrix.
- From the perspective of financial applications, this solution has serious shortcomings. The number of parameters in C is $n(n+1)/2$, which is quadratic in the number of dimensions of the problem. For large numbers of dimensions, the sample covariance matrix may be poorly conditioned (or outright singular).
- We will return to this issue later in these notes.

Kullback-Leibler divergence

- The Kullback-Leibler divergence between two (discrete) probability distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$, where $p_i > 0$, $q_i > 0$ for each i, is defined by

$$\mathrm{KL}(p\|q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. \tag{18}$$

- *First Key Property of the KL divergence.* $\mathrm{KL}(p\|q)$ is convex both in p and in q.
- Indeed, the Hessian $\nabla_p^2 \mathrm{KL}(p\|q)$ is

$$\nabla_p^2 \mathrm{KL}(p\|q) = \begin{pmatrix} 1/p_1 & 0 & \ldots & 0 \\ 0, & 1/p_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1/p_n \end{pmatrix},$$

which is positive definite. The proof of convexity in q is similar.

Kullback-Leibler divergence

- *Second Key Property of the KL divergence.* $\mathrm{KL}(p\|q)$ is nonnegative:

$$\mathrm{KL}(p\|q) \geq 0. \tag{19}$$

- Indeed, consider a minimum of $\mathrm{KL}(p\|q)$. Since the argument p satisfies the constraint $\sum_i p_i = 1$, we are led to the following Lagrange function:

$$L(p, \lambda) = \mathrm{KL}(p\|q) + \lambda\big(\sum_i p_i - 1\big).$$

- Its critical points are given by

$$\begin{gathered}\log \frac{p_i^*}{q_i} + 1 + \lambda^* = 0, \text{ for } i = 1, \ldots, n.\\ \sum_i p_i^* - 1 = 0.\end{gathered}$$

Kullback-Leibler divergence

- Solving this system yields:

$$\begin{aligned}\lambda^* &= -1,\\ p_i^* &= q_i.\end{aligned}$$

- However, $\mathrm{KL}(p^*\|q) = \mathrm{KL}(q\|q) = 0$. Since $\mathrm{KL}(q\|q)$ is convex p^* must correspond to an absolute minimum. Hence, (19) must hold.
- Note that, unlike conventional measures of distance, the KL divergence is, in general, not symmetric in p and q,

$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p). \tag{20}$$

Kullback-Leibler divergence

- There is a close connection between between MLE and KL minimization. The log likelihood function of N observations $y_1, \ldots, y_N$ is

$$\begin{aligned} -\log \mathcal{L}(\theta|y) &= -\sum_{j=1}^{N} \log q(y_j|\theta) \\ &= -\sum_{\text{diff outcomes}} N_j \log q(y_j|\theta) \end{aligned}$$

  where N_j is the count of the outcome y_j.
- This can be rewritten as

$$\begin{aligned} -\frac{1}{N}\log \mathcal{L}(\theta|y) &= -\sum_{j} p_j \log q(y_j|\theta) \\ &= \mathrm{KL}(p\|q) - \sum_{j} p_j \log p_j \\ &= \mathrm{KL}(p\|q) + H(p). \end{aligned}$$

  where $p_j = N_j/N$ is the frequency of the outcome y_j, and $H(p)$ is the entropy of the distribution p.

Regularized approximation

- Typically, estimation problems are *ill posed*, meaning that the number of parameters to be estimated is smaller than the number of data points.
- Ill posed problems are somewhat of a pariah according to classical mathematics (Hadamard), as they do not have unique solutions.
- They are, however, very welcome in data science, as they tend to lead to robust approximate solutions. Stable algorithms for solving ill posed problems typically require using *regularization*.
- The most common regularization is *Tikhonov's regularization*, which is implicitly present in the Levenberg-Marquardt algorithm, and which leads to ridge regression in statistics.

Tikhonov's regularization

- Suppose that we wish to solve the equation
$$Ax = b, \tag{21}$$
that arises in an estimation problem. Here $A \in \mathrm{Mat}_{mn}(\mathbb{R})$ is a square matrix.
- The linear regression approach consists in minimizing the Euclidean norm $\|Ax - b\|_2^2$.
- We may have a preference for a solution which has smaller values x, i.e. we would like to keep $\|x\|_2^2$ small.
- To this end, we consider the objective function in which both terms are present and seek solution to
$$\min \ \|Ax - b\|_2^2 + \lambda\|x\|_2^2. \tag{22}$$
Here, $\lambda > 0$ is the *shrinkage parameter* that measures the trade-off between the two terms.
- This way of modifying an ill posed problem is an example of *Tikhonov's regularization*.

Tikhonov's regularization

- More generally, one may consider the optimization problem
$$\min \ \|Ax - b\|_2^2 + \lambda\|\Gamma x\|_2^2, \tag{23}$$
in which the vector Γx is kept small.
- It has a closed form solution
$$x^* = (A^{\mathrm{T}}A + \lambda\Gamma^{\mathrm{T}}\Gamma)^{-1}A^{\mathrm{T}}y, \tag{24}$$
which is known as *ridge regression*.
- Note how increasing λ impacts the solution by putting more emphasis on the regularization term.
- The value of the parameter λ may be determined by *cross-validation*.

L^1-norm regularization (LASSO)

- Another form of regularization assumes the L^1-norm as the regularization term (a.k.a. LASSO):
$$\min \; \|Ax - b\|_2^2 + \lambda \|x\|_1 . \tag{25}$$
- (25) is a convex quadratic problem (but not strictly convex).
- Unlike ridge regression, the LASSO has no closed form solution.
- An important feature of the LASSO is that, for λ large enough, some components of the optimal x^* are *exactly* zero.
- This is particularly relevant in situations when one believes that some x_i^* should be zero, and seeks a *sparse solution*.
- For this reason, the LASSO is used in machine learning as a *model selection* tool allowing to remove some features among a (possibly large) number of features.

L^1-norm regularization (LASSO)

- Extensions of the L^2 and L^1 regularizations are plentiful.
- The *elastic net regularization* (Hastie and Zou) combines the Tikhonov and the LASSO regularization in a linear fashion:

$$\min \ \|Ax - b\|_2^2 + \lambda_2 \|x\|_2^2 + \lambda_1 \|x\|_1 , \tag{26}$$

  where $\lambda_1, \lambda_2 > 0$.
- The elastic net regularization encourages sparsity but, at the same time, it addresses some of the shortcomings of the pure LASSO model.
- Unlike the LASSO, the elastic net regularization leads to a strictly convex optimization problem.

Covariance matrix estimation with shrinkage

- We return to the issue of estimation of (large) covariance matrix estimation. A popular approach (due to Ledoit and Wolf) consists in finding a linear combination
$$C^* = (1-\alpha)\widehat{C} + \alpha\lambda I \tag{27}$$
of the sample covariance matrix and the identity matrix, which “best” approximates the unknown covariance matrix C.
- In order to formulate the problem we need a measure of distance between matrices,
$$\|A\|_2 = \sqrt{\frac{1}{n}\,\mathrm{tr}(A^{\mathrm{T}}A)} \tag{28}$$
the *Frobenius norm* of the matrix A. It comes associated with the inner product on $\mathrm{Mat}_n(\mathbb{R})$:
$$(A,B) = \frac{1}{n}\,\mathrm{tr}(A^{\mathrm{T}}B) \tag{29}$$
- We would like to solve the problem
$$\min_{\beta,\lambda} \mathsf{E}\big[\|C - ((1-\alpha)\widehat{C} + \alpha\lambda I)\|_2^2\big]. \tag{30}$$

Covariance matrix estimation with shrinkage

- Since C is unknown, this problem cannot be solved. Instead, a representation of the solution can be given that involves some unknown, yet useful, functions.
- We have the identity

$$\mathsf{E}\left[\|C-((1-\alpha)\widehat{C}+\alpha\lambda I)\|_2^2\right]=\alpha^2\|C-\lambda I\|^2+(1-\alpha)^2\mathsf{E}\left[\|C-\widehat{C})\|_2^2\right].$$

- The optimal value for λ is

$$\begin{aligned}\lambda^* &= (C, I)\\ &= \frac{1}{n}\operatorname{tr}(C).\end{aligned}\tag{31}$$

Plugging it back into the objective function above and minimizing over α we find

$$\alpha^*=\frac{\mathsf{E}\left[\|C-\widehat{C}\|_2^2\right]}{\mathsf{E}\left[\|\widehat{C}-\lambda^* I\|_2^2\right]}\,.\tag{32}$$

Covariance matrix estimation with shrinkage

- We claim that $\alpha^* \leq 1$.
- Indeed,

$$\begin{aligned}\mathsf{E}\big[\|\widehat{C}-\lambda^* I\|_2^2\big] &= \mathsf{E}\big[\|\widehat{C}-C+C-\lambda^* I\|_2^2\big]\\ &= \mathsf{E}\big[\|\widehat{C}-C\|_2^2\big]+\|C-\lambda^* I\|_2^2+\big((\widehat{C}-C),(C-\lambda^* I)\big)\\ &\geq \mathsf{E}\big[\|\widehat{C}-C\|_2^2\big],\end{aligned}$$

  since $\big((\widehat{C}-C),(C-\lambda^* I)\big)=0$.
- In summary, we have proved (27) with $\lambda > 0$ and $\alpha < 1$.
- One can set these values *a priori*, or use Monte Carlo simulation.
- In the language of machine learning, the Ledoit-Wolf estimator is the update of the sample covariance $\widehat{C}$ by the shrinkage λI with learning rate α.

Covariance matrix estimation with shrinkage

- Various other approaches, including sparse L^1 regularization, have been proposed.
- Bien and Tibshirani proposed the following approach. Consider the problem:

$$\min\ \operatorname{tr}\big(\widehat{C}C^{-1}\big) + \log\det(C) + \lambda\|\Gamma \circ C\|_1 \quad \text{subject to } C \in \mathbb{P}^n_+,\ C \text{ invertible.}$$

- Here, $\circ$ denotes the Hadamard (elementwise) product of matrices, and $\|A\|_1 = \frac{1}{n}\sum_{i,j}|A_{ij}|$ is the L^1-norm of a matrix.
- Common choices for the matrix Γ are:
  - (i) the matrix of ones, $\Gamma_{ij} = 1$, for all i, j,
  - (ii) this matrix with zeros on the diagonal, $\Gamma_{ii} = 0$, to avoid shrinking the diagonal elements.
- This problem is non-convex, and its solution requires special techniques.
- As usual, the LASSO regularization encourages some entries of C to be exactly zero.

Mean variance optimization

- Consider a portfolio of risky assets $S_1, \ldots, S_n$, and let
  1. (i) r_i denote the return on asset S_i,
  2. (ii) $\mu_i = \mathrm{E}[r_i]$ denote the expected return on S_i,
  3. (iii) $\sigma_i = \sqrt{\mathrm{Var}(r_i)}$ denote the standard deviation of returns on S_i,
  4. (iv) ρ_{ij} denotes the correlation among the returns of S_i and S_j,
  5. (v) $C \in \mathrm{Mat}_n(\mathbb{R})$ denotes the covariance matrix of returns of the assets, $C_{ij} = \rho_{ij}\sigma_i\sigma_j$.
  6. (vi) w_i denotes the weight of asset S_i in the portfolio, $\sum_{i=1}^n w_i = 1$.
- In the Markowitz *mean variance* portfolio problem, we are concerned with the question of allocating the assets in such a way, so that the variance of returns of the portfolio returns is minimal, while the expected return is at least a certain target level r.
- Additionally, one imposes other inequality and equality constraints which reflect the portfolio manager's (PM) mandate and views. For example, if the portfolio is long only, part of the inequality constraints will read $w_i \geq 0$.

Mean variance optimization

- In other words, the Markowitz *mean variance* optimization problem is formulated as the following convex optimization problem:

$$\min \ \sigma^2 = \frac{1}{2} w^{\mathrm{T}} C w \quad \text{subject to} \begin{cases} \mu^{\mathrm{T}} w \geq r, \\ Aw = b, \\ Bw \geq c. \end{cases} \tag{33}$$

- Let $\lambda_r, \lambda_{\mathcal{E}}, \lambda_{\mathcal{I}}$ denote the Lagrange multipliers corresponding to the three constraints above. The Lagrange function reads:

$$L(w, \lambda_r, \lambda_{\mathcal{E}}, \lambda_{\mathcal{I}}) = \frac{1}{2} w^{\mathrm{T}} C w - \lambda_r(\mu^{\mathrm{T}} w - r) - \lambda_{\mathcal{E}}^{\mathrm{T}}(Aw - b) - \lambda_{\mathcal{I}}^{\mathrm{T}}(Bw - c). \tag{34}$$

Mean variance optimization

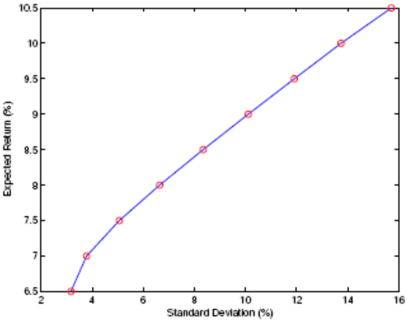
- The first order KKT conditions read:

$$\begin{aligned} Cw &= \lambda_r \mu + A^{\mathrm{T}} \lambda_{\mathcal{E}} + B^{\mathrm{T}} \lambda_{\mathcal{I}}, \\ \mu^{\mathrm{T}} w &\geq r, \\ Aw &= b, \\ Bw &\geq c, \\ \lambda_r(\mu^{\mathrm{T}} w - r) &= 0, \\ \lambda_{\mathcal{I}}(Bw - c) &= 0, \\ \lambda_r &\geq 0, \\ \lambda_{\mathcal{I}} &\geq 0. \end{aligned} \tag{35}$$

- Whether these conditions have a solution or not, and what is the optimal value, depends on the values of the parameters in the constraints.

Mean variance optimization

- In particular, there is a range $r \in [r_{min}, r_{max}]$ of target expected portfolio returns, for which the problem has a solution.
- The curve representing the expected return r as a function of the optimal standard deviation σ of portfolio returns is called the *efficient frontier*.

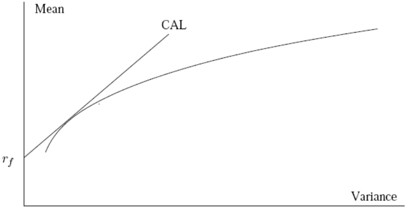


Mean variance optimization

- It is apparent the solution of the optimality conditions (35) involves inverting the covariance matrix C.
- For this reason, it is of utmost importance that the estimated covariance matrix is well conditioned and has a well behaved inverse. As explained earlier in these notes, a great deal of effort has been made to develop reliable methodologies for covariance matrix estimation, especially for large portfolios.
- Without additional measures, Markowitz MV optimization typically leads to highly unintuitive, extreme portfolios, in which outsized long positions in some assets are offset by outsized short positions in other assets. Such portfolios, even though mathematically feasible are practically totally unrealistic.
- Frequently (and IMHO incorrectly) this phenomenon is attributed to estimation errors of the “true” covariance matrix. These errors are supposed to be remediated by the regularization techniques discussed earlier.

Maximizing the Sharpe ratio

- Assume now that the portfolio contains a riskless asset with deterministic rate of return is r_f. It is natural to expect that $r_f < r$.
- We allocate a fraction $0 \leq \eta \leq 1$ of of the portfolio into the riskless asset. Risk / return profiles for different values of η can be represented as a straight line a capital allocation line (CAL) on the standard deviation / mean plane.



- The optimal CAL lies below all the other CALs with $r > r_f$, since the corresponding portfolios have the lowest σ for any given value of $r > r_f$.

Maximizing the Sharpe ratio

- It follows that the optimal CAL goes through a point on the efficient frontier and never goes above a point on the efficient frontier.
- In other words, the slope of the optimal CAL is the derivative of the function $r(\sigma)$ that defines the efficient frontier, and so *the optimal CAL is tangent to the efficient frontier*.
- The point where the optimal CAL touches the efficient frontier corresponds to the optimal risky portfolio.

Maximizing the Sharpe ratio

- Alternatively, one can think of the optimal CAL as the one with the smallest slope. This is the portfolio that maximizes the *Sharpe ratio*:

$$\mathrm{SR}(w) = \frac{\mu^{\mathrm{T}} w - r_f}{\sqrt{w^{\mathrm{T}} C w}}\,. \tag{36}$$

- The corresponding optimization problem

$$\max\ \frac{\mu^{\mathrm{T}} w - r_f}{\sqrt{w^{\mathrm{T}} C w}} \quad \text{subject to} \begin{cases} Aw = b, \\ Bw \geq c \end{cases} \tag{37}$$

  is not concave, and it is hard to solve.
- It can be replaced with an equivalent convex quadratic problem as follows.

Maximizing the Sharpe ratio

- The feasible set of (37), $\mathcal{P} = \{w \in \mathbb{R}^n : \ Aw = b, Bw \geq c\}$, is a polyhedron. Define the convex set

$$\mathcal{P}^+ = \{(y, \kappa) \in \mathbb{R}^{n+1} : \ \kappa > 0, \text{ and } y/\kappa \in \mathcal{P}\} \cup (0, 0).$$

- Then the portfolio w^* optimizing (37) is of the form $w^* = y^* / \kappa^*$, where the pair (y^*, κ^*) is the solution to the following convex quadratic problem:

$$\min \ y^{\mathrm{T}} C y \quad \text{subject to} \begin{cases} (y, \kappa) \in \mathcal{P}^+, \\ (\mu - r_f)^{\mathrm{T}} y = 1. \end{cases} \tag{38}$$

- In order to the the equivalence of the two problems, we substitute in (37):

$$\kappa = \frac{1}{(\mu - r_f)^{\mathrm{T}} y},$$
$$y = \kappa w.$$

Benchmark tracking

- In most quantitative asset management environments, portfolios are selected with respect to a particular benchmark in mind. Typically, the benchmark is a market index.
- The PM's mandate may be, for example, to closely track the benchmark, or outperform it by a certain amount.
- The *tracking error* (or *excess return*) of a portfolio with a given benchmark is the difference between the returns of the portfolio and the benchmark. It is defined as:
$$r^{\mathrm{T}}w - r^{\mathrm{T}}w_{bench} = r^{\mathrm{T}}(w - w_{bench}), \tag{39}$$
  where r^{T} is the vector of returns of the assets and w_{bench} is the vector of weights of the assets in the benchmark.
- The *ex ante* (predicted) tracking error is defined as
$$\epsilon(w) = \sqrt{(w - w_{bench})^{\mathrm{T}}C(w - w_{bench})} \tag{40}$$

Benchmark tracking

- The PM whose mandate is to track the benchmark with maximum ex ante tracking error ϵ faces the following convex optimization problem:

$$\max_w r^{\mathrm{T}}(w - w_{bench}) \quad \text{subject to} \begin{cases} w^{\mathrm{T}} C w \leq \sigma^2, \\ (w - w_{bench})^{\mathrm{T}} C (w - w_{bench}) \leq \epsilon^2, \\ Aw = b, \\ Bw \geq c. \end{cases} \tag{41}$$

- Unlike the Markowitz problem (33) that has linear constraints only, it is not in standard quadratic programming (the constraint limiting the portfolio tracking error is quadratic), which makes it harder to solve.
- The tracking error constraint is, however, a convex quadratic function and, as discussed in Lecture Notes #4, we can rewrite this constraint in conic form. The resulting problem is a second-order cone optimization problem.

Benchmark tracking

- Since C is positive definite, there is a non-singular $R \in \mathrm{Mat}_n(\mathbb{R})$ such that $C = RR^{\mathrm{T}}$. We introduce the variables:

$$\begin{aligned} y_0 &= \sigma, \\ y &= R^{\mathrm{T}} w, \\ z_0 &= \epsilon, \\ z &= R^{\mathrm{T}}(w - w_{bench}). \end{aligned} \tag{42}$$

- With these definitions, the first two constraints in (41) say that the points $(y_0, y) \in \mathbb{R}^{n+1}$ and $(z_0, z) \in \mathbb{R}^{n+1}$ are elements of the second-order cone $\mathcal{K}_n$ in $\mathbb{R}^{n+1}$.

Benchmark tracking

- The benchmark tracking problem can be formulated as the following SOCP:

$$\max_{w} r^{\mathrm{T}}(w - w_{bench}) \quad \text{subject to} \begin{cases} Aw = b, \\ Bw \geq c, \\ R^{\mathrm{T}} w - y = 0 \\ R^{\mathrm{T}} w - z = R^{\mathrm{T}} w_{bench} \\ y_0 = \sigma, \\ z_0 = \epsilon, \\ (y_0, y) \in C_n, \\ (z_0, z) \in C_n. \end{cases} \tag{43}$$

References

Boyd, S., and Vanderberghe, L.: *Convex Optimization*, Cambridge University Press (2004).

Cornuejos, G., and Tutuncu, R.: *Optimization Methods in Finance*, Cambridge University Press (2007).