Portfolio Optimization in a Big Data Context

Thierry Bazier-Matte

Summer 2015

Notation. In the following, $\boldsymbol{A}$ (capital boldface) are assumed to represent a real subset of any dimension, A (capital case) represents random variables (or distributions) and a (lower case) represents deterministic variables or realizations. $\mathscr{R}$ represents the real set.

Let $M = (X, R)$ the *market* be an unknown distribution with support $\boldsymbol{M} = \boldsymbol{X} \times \boldsymbol{R} \subseteq \mathscr{R}^{p+1}$, ie. numerically qualifiable, with $(x, r) = m \sim M$ a *market observation*, consisting in one part *state* $x \in \mathscr{R}^p$ and another part *outcome* $r \in \mathscr{R}$. Typically x is a vector of observations from various variable of interests, such as financial or economical news, etc. Scalar r in this article shall represent the return from a financial asset of interest. Finally, let $M_n = \{M, \ldots, M\}$ be a *random set* of n (unrealized) observations (with support $\boldsymbol{M}^n$). Therefore $\mu_n \sim M_n$ represents an iid sample of n market observations.

This article studies *linear investment decisions* $q^T x$, with $q \in \boldsymbol{Q} \subseteq \mathscr{R}^p$.

Assumption. We suppose that observed returns r are constrained by $|r| \leq \bar{r}$ with probability $1 - \delta_r$ and that observed states x are constrained by $\|x\|_2 \leq X_{\max}$ with probability $1 - \delta_x$.

Definition. Let $\ell : \boldsymbol{M} \times \boldsymbol{Q} \to \mathscr{R}$ be a *loss function* defined by

$$\ell(m, q) = \ell(x, r, q) = -u(r\, q^T x + R_f(1 - q^T x)),$$

where R_f is the risk free return rate and $u(r) = \min(r, \beta r)$, with $0 < \beta < 1$ the risk aversion parameter. We also define the *cost function* $c : \mathscr{R} \times \boldsymbol{R} \to \mathscr{R}$ as

$$c(p, r) = -u(pr + (1 - p)R_f),$$

so that $\ell(x, r, q) = c(q^T x, r)$.

Definition. The *empirical risk* $\hat{R} : \boldsymbol{M}^n \times \boldsymbol{Q} \to \mathscr{R}$ associated with decision q and market sample μ_n is given by

$$\hat{R}_{\mu_n}(q) = n^{-1} \sum_{i=1}^{n} \ell(m_i, q).$$

Definition. The *empirical decision algorithm* $\hat{A}_n : \boldsymbol{M}^n \to \boldsymbol{Q}$ associated with market sample μ_n is the optimal value of the problem

$$\text{minimize} \quad \hat{R}_{\mu_n}(q) + \lambda\|q\|_2^2.$$

From now on, $\hat{q}_n := \hat{A}_n(\mu_n)$ the empirical decision associated with market sample μ_n and $\hat{Q}_n := A_n(S_n)$ the random empirical decision, ie. $\hat{q}_n \sim \hat{Q}_n$.

Definition. The *true risk* $R_{\text{true}} : \boldsymbol{Q} \to \mathscr{R}$ associated with decision q is given by

$$R_{\text{true}}(q) = E_M[\ell(m, q)].$$

Definition. The *optimal decision* $q^\star$ is the optimal value of the problem

$$\text{minimize} \quad R_{\text{true}}(q) + \lambda\|q\|_2^2.$$

1 Stability Definitions and Theorems

Definition. Let $\hat{q}_n = \hat{A}_n(\mu_n)$ and $\hat{q}_{n\setminus i} = \hat{A}_n(\mu_{n\setminus i})$, where μ_n and $\mu_{n\setminus i}$ only differs in their i^{th} observation, which has been redrawn from M in the case of $\mu_{n\setminus i}$. The algorithm $\hat{A}_n$ is said to have *uniform stability* α_n if, for any $m \sim M$,

$$|\ell(m, \hat{q}_n) - \ell(m, \hat{q}_{n\setminus i})| \leq \alpha_n.$$

Definition. A loss function ℓ is σ-admissible if its cost function c is convex with respect to p the investment decision and the following holds for any p_1, p_2 and r:

$$|c(p_1, r) - c(p_2, r)| \leq \sigma|p_1 - p_2|.$$

Remark. The loss function as defined above is σ-admissible with $\sigma = \bar{r} + R_f$.

Theorem 1. *If ℓ is σ-admissible and if, for any $x \in \boldsymbol{X}$, $\|x\|_2^2 \leq X_{\max}^2$, then $\hat{A}_n$ has uniform stability with*

$$\alpha_n = \frac{\sigma^2 X_{\max}^2}{2\lambda n}.$$

Proof. See Bousquet, Theorem 22. □

We therefore conclude that $\hat{A}_n$ has uniform stability with

$$\alpha_n = \frac{(\bar{r} + R_f)^2 X_{\max}^2}{2\lambda n}.$$

Theorem 2. *If $\hat{A}_n$ has uniform stability α_n and the loss function is such that for any $m \sim M$ and any $\hat{q}_n = \hat{A}_n(\mu_n)$, $0 \leq \ell(m, \hat{q}_n) \leq B_n$, then for any $\delta \in (0,1)$, the following bound holds with probability at least $1 - \delta$ over the random sample draw $\mu_n \sim M_n$:*

$$|R_{\text{true}}(\hat{q}_n) - \hat{R}(\hat{q}_n)| \leq 2\alpha_n + (4n\alpha_n + B_n)\sqrt{\frac{\log(2/\delta)}{2n}}.$$