STABLE LINEAR-TIME OPTIMIZATION IN ARBITRAGE PRICING THEORY MODELS

GORDON RITTER*

Courant Institute of Mathematical Sciences
New York University
251 Mercer St., New York, NY 10012

ABSTRACT. We present an explicit formula for mean-variance optimization in the context of APT models (also called multi-factor models), and related generalizations with trading costs. Our explicit formula has two desirable features:

(1) the solutions are well-defined and numerically stable in the presence of approximate or exact colinearity in the design matrix, and
(2) the computational complexity is (manifestly) linear with respect to the number of assets.

1. Markowitz Optimization and APT

1.1. **APT.** Many models for asset returns in empirical finance, following Ross (1976), assume a linear functional form

$$R_{t+1} = X_t f_t + \epsilon_t, \quad \mathbb{E}[\epsilon] = 0, \quad \mathbb{V}[\epsilon] = D \tag{1}$$

where R_{t+1} is an n-dimensional random vector containing the cross-section of returns in excess of the risk-free rate over some time interval $[t, t+1]$, X_t is a (non-random) $n \times p$ matrix that is known before time t, and ϵ_t is assumed to follow a mean-zero distribution with diagonal variance-covariance matrix

$$D := \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2) \text{ with all } \sigma_i^2 > 0. \tag{2}$$

Eq. (2) entails that all significant sources of correlation are already captured by factors, represented as columns of X_t. We henceforth suppress the implicit time index since most of our discussions concern a single time interval.

The variable f in (1) denotes a p-dimensional random vector process which cannot be observed directly; information about the f-process must be obtained via statistical inference. Statistically estimated realizations $\hat{f}$ are

Date: December 2, 2017.
* *E-mail address:* ritter@post.harvard.edu.

called *factor returns* by practitioners (Menchero et al., 2008). We assume that the f-process has finite first and second moments given by

$$\mathbb{E}[f] = \mu_f, \quad \text{and} \quad \mathbb{V}[f] = F. \tag{3}$$

The model (1), (2) and (3) entails associated reductions of the first and second moments of the asset returns:

$$\mathbb{E}[R] = X\mu_f, \quad \text{and} \quad \Sigma := \mathbb{V}[R] = D + XFX' \tag{4}$$

where X' denotes the transpose. To use (4) in portfolio construction, estimates of parameters (3) must be obtained by statistical inference, as discussed in the next section.

The functional form (1) is general enough to include the three-factor model of Fama and French (1993), a wide class of models compatible with the arbitrage pricing theory (APT) of Ross (1976) and Roll and Ross (1980), the partially-predictable return-generating process of Gârleanu and Pedersen (2013), and others. Models of the form (1) are also used at most large banks and asset management firms, often as risk models. Connor et al. (2010), Fabozzi et al. (2010), and Menchero et al. (2008) give examples of factor models used in practice.

1.2. **Identifiability.** The model (1) is said to be *identifiable* if any of the following equivalent conditions hold:

(a) $\operatorname{rank}(X) = p$, ie. X is full rank,
(b) $X'X$ is invertible, where X' denotes the transpose
(c) The function $\ell(f) = \|R - Xf\|^2$ has a unique minimizer.
(d) The ordinary least-squares (OLS) estimator of f exists.

We call a model *barely identifiable* (or approximately collinear) if conditions (b) or (c) are close to being violated; e.g. if $X'X$ has a very small eigenvalue, or equivalently, if $\ell(f)$ has a direction of near-zero curvature, or the Hessian of $\ell(f)$ is nearly degenerate. Identifiability might seem an important condition to require when building a factor model, but, in practice, unidentifiable and barely identifiable models arise naturally and often.

Example 1. Consider a unified model for the European equity market. Significant drivers of asset return covariance include market beta, industry membership, country membership, and others. Define

$$N_{i,j} = \begin{cases} 1, & \text{stock } i \text{ is a member of industry } j \\ 0 & \text{otherwise} \end{cases}$$

with C defined similarly for countries. Let B denote an $n \times 1$ column vector of market betas. The full design matrix is $X = \begin{bmatrix} B & C & N \end{bmatrix}$ and is of course not identifiable.[1]

Example 2. Suppose we augment model (1) with two closely-related alpha forecasts, which we call α_1 and α_2 (such as earnings yield with two different types of earnings). Collect these into an $n \times 2$ matrix $A = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix}$. The augmented design matrix is then

$$X = \begin{bmatrix} A & B & C & N \end{bmatrix}$$

We now have an approximate and an exact colinearity in the same model!

Our intended application is to models which generalize Example 2, and hence fall outside of strong-form market efficiency by allowing for the existence of alpha factors. Many authors in in empirical finance either propose examples of alpha factors (Asness et al., 2013) or discuss optimization in a context where some alpha factors are given (Gârleanu and Pedersen, 2013). This paper is of the latter type.

Identifiability is not necessary for the model (1) to be a correct description of the world. Referring to Example 1, there can, of course exist latent, unobservable stochastic processes f_j corresponding to industries and countries which drive returns according to eq. (1). However, lack of identifiability complicates estimation of $\hat{f}$ and by extension, μ_f and F.

If the model is identifiable, then reasonable estimates for the factor returns are

$$\hat{f}_{OLS} = (X'X)^{-1}X'R,$$

and μ_f can be estimated by the time-series mean of $\hat{f}_{OLS}$. In the unidentifiable case, $\hat{f}_{OLS}$ doesn't exist, and in the barely identifiable case, $\hat{f}_{OLS}$ misleadingly contains large opposing coefficients for strongly-correlated factor pairs.

Several well-known methods exist for obtaining estimates of $\hat{f}$ in the unidentifiable case. Intuitively, when there are multiple solutions to the first-order condition, some constraint or prior must be introduced to prefer one solution over another. One may remove exact colinearities by imposing $p - \text{rank}(X)$ linear constraints on the coefficients. This *restricted least squares* method (Greene and Seaks, 1991) is simple and classical, but ultimately incomplete. It would handle Example 1 but not Example 2, and the constraints must be decided arbitrarily. A more general class of inference

[1]An exact colinearity, such as in Example 1, can be rectified by choosing a basis for the column space of X and replacing X with a (thinner) matrix having the elements of this basis as columns. This is less than ideal for two reasons: it loses the economic meanings of the factors, and doesn't easily generalize to the case of approximate colinearity, ie. barely identifiable models.

procedures are provided by Bayesian regression, including the popular ridge, lasso, and elastic-net estimators as special cases; see Zou and Hastie (2005). Such procedures have the added benefit of providing an explicit model complexity parameter to be used in cross-validation; see Friedman et al. (2001) for details.

In particular, the *ridge* estimator is

$$\hat{f}_\lambda = \arg\min_f \left[\|R - Xf\|^2 + \lambda\|f\|^2\right] = (X'X + \lambda I)^{-1}X'R, \quad \lambda > 0. \tag{5}$$

For any $\lambda > 0$ and any real matrix X, $\hat{f}_\lambda$ exists. Moreover, the limit

$$\lim_{\lambda\to 0} \hat{f}_\lambda = X^+R \tag{6}$$

also exists, where X^+ denotes the Moore-penrose pseudoinverse of X; see Albert (1972) for background. By a classical result, (6) is the minimum-norm vector among all minimizers of the least-squares objective, and gives yet another way of obtaining coefficient estimates in the unidentifiable case.

The above procedures provide a reasonable guide for coefficient estimation and, by extension, to statistical inference and estimation of μ_f and F in unidentifiable models. However our goals extend beyond coefficient estimation to portfolio optimization and hedging. To this end, we present an explicit formula for Markowitz (1952) mean-variance optimization, and extensions with trading costs, which has two important features:

(1) the solutions are well-defined and numerically stable in the presence of approximate or exact colinearity in the design matrix, and
(2) the computational complexity is (manifestly) linear with respect to the number of assets.

Our solution could be called "optimization in factor space," meaning that it first finds the risk and alpha factor exposures of the optimal portfolio, before representing the optimal portfolio as a linear transformation of the optimal exposures.

2. Optimization

The Markowitz (1952) mean-variance problem with moments (4) is

$$h^* = \operatorname{argmax} f(h) \quad \text{where} \tag{7}$$

$$f(h) = h'X\mu_f - \frac{\kappa}{2}h'XFX'h - \frac{\kappa}{2}h'Dh \tag{8}$$

and where $\kappa > 0$ is the Arrow-Pratt constant absolute risk aversion. The first two terms in (7) depend on h only through its *exposures*, defined by

$$q := X'h \tag{9}$$

In terms of (9) the first two terms in (7) can be written more simply as

$$q'\mu_f - (\kappa/2)q'Fq$$

Industry-standard terminology for the third term, $h'Dh$, is *idiosyncratic variance*. A key insight is that at optimality, the third term can be written as a function of q as well:

Intuition 1. *An optimal portfolio h for* (7) *must minimize idiosyncratic variance $h'Dh$ among all portfolios with the same exposures $q = X'h$.*

With this intuition in mind to clarify the proof, we can now proceed to the main result of the paper.

Theorem 1. *The risk/alpha exposures of the portfolio optimizing* (7) *are q^* and the optimal holdings are h^*, where*

$$q^* = \kappa^{-1}[F + [X'D^{-1}X]^+]^{-1}\mu_f \tag{10}$$

$$h^* = \kappa^{-1}D^{-1/2}(X'D^{-1/2})^+\left[F + (X'D^{-1}X)^+\right]^{-1}\mu_f \tag{11}$$

Proof. Let $h^*(q)$ be the solution to

$$h^*(q) = \text{argmin}_h\, h'Dh \quad \text{subject to:} \quad X'h = q. \tag{12}$$

Let

$$V(q) = h^*(q)'Dh^*(q)$$

be the minimum idiosyncratic variance (still subject to $X'h = q$). Using Intuition 1, the mean-variance objective can then be written entirely in terms of q:

$$\max_q \left\{ q \cdot \mu_f - \frac{\kappa}{2} q'Fq - \frac{\kappa}{2} V(q) \right\}. \tag{13}$$

Finding $V(q)$ will allow us to directly attack (13), hence we now devote ourselves to this task. We show in due course that $V(q)$ is quadratic and can be written down explicitly.

Changing variables to $\eta := D^{1/2}h$ the problem (12) is

$$\min_\eta \|\eta\|^2 \text{ subject to } X'D^{-1/2}\eta = q \tag{14}$$

Eq. (14) is a special case of a classical problem which has a beautiful solution and deep connections to other areas of mathematics. The minimum norm solution of a linear system $Bx = q$ is given by $x = B^+q$ where B^+ is the *pseudoinverse* of B, in the sense of Moore (1920) and Penrose (1955).

Computation of B^+ is straightforward given the singular value decomposition: if $B = U\Sigma V'$ where U and V are orthogonal and Σ is rectangular-diagonal, then

$$B^+ = V\Sigma^+U' \tag{15}$$

where Σ^+ is computed by inverting the non-zero diagonal elements and taking the transpose.[2]

Hence the solution to (14) is given by[3]

$$\eta^* = (X'D^{-1/2})^+ q \quad \Rightarrow \quad h^*(q) = D^{-1/2}(X'D^{-1/2})^+ q \tag{16}$$

and therefore

$$\begin{aligned} V(q) &= h^*(q)' D h^*(q) \\ &= [D^{-1/2}(X'D^{-1/2})^+ q]' D [D^{-1/2}(X'D^{-1/2})^+ q] \\ &= q'[X'D^{-1}X]^+ q \end{aligned}$$

The third term in (13) can thus be combined with the second term in (13) to form a single quadratic term. The optimal q is then given by

$$q^* = \kappa^{-1}[F + [X'D^{-1}X]^+]^{-1}\mu_f$$

and the optimal holdings h^* are found by plugging q^* into (16). □

Theorem 2. *Let the number of factors, p, be a fixed constant. The computational complexity of finding the optimal exposures and holdings, Eqns.* (10)–(11)*, is linear-time in n, the number of assets.*

Proof. Let $p \ll n$; the "economical" SVD of an $n \times p$ matrix can be computed (Golub and Van Loan, 2012) in about

$$6np^2 + 20p^3 \quad \text{flops} \tag{17}$$

This bounds the complexity of the pseudoinverse and actual inverse required in (11), since the Moore-Penrose pseudoinverse is given in terms of the SVD by (15). But if p is constant then (17) is linear in n. □

We now make several observations concerning the above which may be useful to practitioners. We lose no generality in assuming that the outputs from m distinct alpha models are stored in the first m columns in X, and defining $k = p - m$ as the number of risk factors, one has

$$X = \begin{bmatrix} X_\alpha & X_\sigma \end{bmatrix} \in \mathbb{R}^{n \times (k+m)} \tag{18}$$

Several of the remarks below refer to the notation of (18).

[2]This provides an automatic way to ensure numerical stability: computing Σ^+ is nothing more than computing the multiplicative inverses of a sequence of real numbers; one should treat those numbers as zero if they are within floating-point precision! More aggressive regularization can be obtained by further increasing this threshold away from the floating-point "epsilon."

[3]Under certain conditions on matrices A and B, one has $(AB)^+ = B^+A^+$ but none of those conditions apply here, so the right-hand side of (16) can't be simplified further.

Remark 1. As stated above we do not assume X is of full rank. Thus our method deals gracefully with approximate or exact colinearities among the risk factors and alpha factors (or X_σ and X_α). This could arise if X_σ contains indicator variables for two classifications, such as sector and country, if two or more alpha models were closely related representations of the same model/dataset, or if some group of alpha factors were approximately spanned by the risk factors. Eqns. (10)–(11) remain valid if there *aren't* any collinearities of course, so this approach allows one formula to cover all cases.

Remark 2. The number of risk factors k, the number of alpha models m, and hence the overall number of factors $p = k + m$ is ultimately a modeling choice, but since the complexity scales as p^3 for fixed n, parsimonious models are more efficiently optimized. Parsimonious models are also preferred in statistical model selection procedures according to the Ockham's razor principle (Jefferys and Berger, 1992). Accordingly, one can select the number of factors (or model complexity) by splitting the full data into a training set, and a testing (or out-of-sample) set, and setting the model complexity via cross-validation within the training set. A full description of the procedure is beyond our current scope, but an excellent treatment can be found in Friedman et al. (2001, Chapter 7).

Remark 3. The technique above can be extended to include certain simple trading cost models. For example, if we have a starting portfolio h_0 and quadratic trading costs[4] given by

$$(h - h_0)'\Lambda(h - h_0) \quad \text{where} \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n). \tag{19}$$

then (7) becomes (with the shorthand $\alpha := X\mu_f$)

$$f(h) = h'\alpha - \frac{\kappa}{2}h'XFX'h - \frac{\kappa}{2}h'Dh - (h - h_0)'\Lambda(h - h_0) \tag{20}$$

$$= h'(\alpha + 2\Lambda h_0) - \frac{\kappa}{2}h'XFX'h - \frac{\kappa}{2}h'(D + \frac{2}{\kappa}\Lambda)h \tag{21}$$

The latter has the same mathematical structure as the original problem (7), so Theorem 1 applies. Even if trading costs are not of the form (19), it is still useful to be able to compute the Markowitz portfolio in linear time, because the solution to the more complicated problem amounts to tracking the Markowitz portfolio in a cost-efficient manner (Kolm and Ritter, 2015).

Remark 4. Gârleanu and Pedersen (2013) show that under certain assumptions on return predictability and trading costs, the dynamically optimal

[4]This trading cost model is too simple to be used in practice because the quadratic structure tends to underestimate the cost of small trades.

portfolio sequence is given by a linear combination of past optimal portfolios and the "aim portfolio." Aim portfolios are weighted sums of future Markowitz portfolios over many horizons:

$$\text{aim}_t = \sum_{\tau=t}^{\infty} z(1-z)^{t-\tau}\mathbb{E}_t[\text{Markowitz}_\tau] \tag{22}$$

where Markowitz_τ is Gârleanu and Pedersen (2013) notation for what we call h^*, a solution to the problem (7). Theorem 1 and 2 imply that the aim portfolio (22), after suitable truncation of the infinite sum, can be computed with $O(n)$ efficiency and stably in the presence of colinearities, ie. that the computations implied by (22) are feasible even for large n.

Remark 5. If the alpha factors are statistically independent from the risk factors, then F must have a block structure with blocks F_α and F_σ. A portfolio *neutral* to all of the columns of X_σ (the risk factors), may be obtained as the limit of h^* as $[F_\sigma]_{i,i} \to +\infty$ for all $i = 1, \ldots, k$. This limit exists, and is equivalent to solving the optimization with the k linear constraints $h'X_\sigma = 0$. Under the independence assumption of alpha factors and risk factors, the constrained factor-neutral portfolio will usually be close to h^*, but not exactly the same since the constrained solution may have higher idiosyncratic variance than other unconstrained solutions.

3. Conclusions

Theorem 1 is, primarily, a research tool, useful whenever "factor returns" or "factor-mimicking portfolios" (Connor et al., 2010; Menchero et al., 2008, for details) are useful. It can be generalized to include very simple transaction cost models, and is also useful as an input to multiperiod frameworks such as Gârleanu and Pedersen (2013). Our formula can be computed very efficiently for large portfolios driven by a small number of factors, ie. when $p \ll n$. It exhibits the optimal portfolio as a linear transformation of the optimal factor exposures, and hopefully clarifies how Ross (1976) pricing theory is the key to using Markowitz (1952) optimization when the number of assets, n, is large.

References

Albert, A. (1972). Regression and the moore-penrose pseudoinverse. Technical report.

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, 68(3):929–985.

Connor, G., Goldberg, L. R., and Korajczyk, R. A. (2010). *Portfolio risk analysis*. Princeton University Press.

Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2010). *Quantitative equity investing: Techniques and strategies*. John Wiley & Sons.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Gârleanu, N. and Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance*, 68(6):2309–2340.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Greene, W. H. and Seaks, T. G. (1991). The restricted least squares estimator: a pedagogical note. *The Review of Economics and Statistics*, pages 563–567.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72.

Kolm, P. N. and Ritter, G. (2015). Multiperiod portfolio selection and bayesian dynamic models. *Risk*.

Markowitz, H. (1952). Portfolio selection*. *The journal of finance*, 7(1):77–91.

Menchero, J., Morozov, A., and Shepard, P. (2008). The barra global equity model (gem2). *MSCI Barra Research Notes*, page 53.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press.

Roll, R. and Ross, S. A. (1980). An empirical investigation of the arbitrage pricing theory. *The Journal of Finance*, 35(5):1073–1103.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of economic theory*, 13(3):341–360.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.