# Notes by Daniel

December 22, 2025

## 1 Core Problem Formulation

Consider a symmetric positive semi-definite matrix $A \in \mathbb{R}^{p\times p}$ and a vector $\boldsymbol{b} \in \mathbb{R}^p$. For each binary vector $\boldsymbol{s} \in \{0,1\}^p$, let $A_{[\boldsymbol{s}]}$ denote the principal submatrix of $A$ indexed by $\{j : s_j = 1\}$, and $\boldsymbol{b}_{[\boldsymbol{s}]}$ as the subvector of $\boldsymbol{b}$ on the same index set. For a given positive integer $k \leq p$ (usually $k$ is much smaller than $p$), our aim is solve the sparse constrained problem:

$$\max_{\boldsymbol{s}\in\{0,1\}^p} \boldsymbol{b}_{[\boldsymbol{s}]}^\top (A_{[\boldsymbol{s}]})^\dagger \boldsymbol{b}_{[\boldsymbol{s}]}, \quad \text{subject to } \ |\boldsymbol{s}| \leq k. \tag{1}$$

where $|\boldsymbol{s}|$ denotes the number of ones in $\boldsymbol{s}$ and $(A_{[\boldsymbol{s}]})^\dagger$ denotes the Moore-Penrose pseudo-inverse of $A_{[\boldsymbol{s}]}$. It is important to keep in mind that $(A_{[\boldsymbol{s}]})^\dagger$ is not a submatrix of $A^\dagger$. Note that if $A$ is full-rank (i.e., positive definite), each $A_{[\boldsymbol{s}]}$ is invertible. Also note that each $A_{[\boldsymbol{s}]}$ is (symmetric) positive semi-definite since it is a principal submatrix of $A$, which is assumed to be symmetric positive semi-definite. The problem (**??**) is NP-hard and serves as a unifying framework for several important problems in statistics and machine learning. Below we detail how various domain-specific problems can be reformulated as instances of this core problem. We now list some of these problems.

**Example 1.** *(***Minimum-Variance Portfolio Selection:***)* The classical minimum-variance portfolio optimization typically produces dense solutions with nonzero weights across all assets, which can lead to high transaction costs and increased estimation error. To address these limitations, sparse portfolio selection incorporates an explicit cardinality constraint that limits the number of assets to at most $k$. This yields the optimization problem gievn by

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}, \quad \text{subject to } \mathbf{1}^\top \boldsymbol{\beta} = 1, \ \|\boldsymbol{\beta}\|_0 \leq k, \tag{2}$$

where $\Sigma \in \mathbb{R}^{p\times p}$ is an invertible covariance matrix of asset returns. This formulation can be equivalently expressed as a binary-constrained optimization problem:

$$\min_{\substack{\boldsymbol{s}\in\{0,1\}^p \\ |\boldsymbol{s}|\leq k}} \ \min_{\substack{\boldsymbol{\beta}_{[\boldsymbol{s}]}\in\mathbb{R}^{|\boldsymbol{s}|} \\ \mathbf{1}^\top\boldsymbol{\beta}_{[\boldsymbol{s}]}=1}} \boldsymbol{\beta}_{[\boldsymbol{s}]}^\top \Sigma_{[\boldsymbol{s}]} \boldsymbol{\beta}_{[\boldsymbol{s}]}. \tag{3}$$

For any fixed support $\boldsymbol{s}$, the inner minimization admits a closed-form solution, given by

$$\boldsymbol{\beta}^*_{[\boldsymbol{s}]} = \frac{(\Sigma_{[\boldsymbol{s}]})^{-1}\mathbf{1}}{\mathbf{1}^\top(\Sigma_{[\boldsymbol{s}]})^{-1}\mathbf{1}},$$

yielding the minimal portfolio variance $1/\mathbf{1}^\top(\Sigma_{[\boldsymbol{s}]})^{-1}\mathbf{1}$. Thus, the combinatorial optimization problem reduces to

$$\max_{\boldsymbol{s}\in\{0,1\}^p} \mathbf{1}^\top(\Sigma_{[\boldsymbol{s}]})^{-1}\mathbf{1}, \quad \text{subject to } |\boldsymbol{s}| \leq k. \tag{4}$$

This precisely matches the form of the core problem (**??**) with $A = \Sigma$ and $\boldsymbol{b} = \mathbf{1}$, demonstrating that sparse portfolio selection is another important instance of the core problem.

**A Crucial Observation:** Recall the core problem (**??**). Assume that $A$ has a full rank (i.e., positive definite) and all the elements of $\boldsymbol{b}$ are non-zero. Then, for every $\boldsymbol{s} \in \{0,1\}^p$,

$$\begin{aligned}
\boldsymbol{b}_{[\boldsymbol{s}]}^\top (A_{[\boldsymbol{s}]})^\dagger \boldsymbol{b}_{[\boldsymbol{s}]} &= \boldsymbol{b}_{[\boldsymbol{s}]}^\top (A_{[\boldsymbol{s}]})^{-1} \boldsymbol{b}_{[\boldsymbol{s}]} \\
&= \boldsymbol{b}_{[\boldsymbol{s}]}^\top (A_{[\boldsymbol{s}]})^{-1} \boldsymbol{b}_{[\boldsymbol{s}]} \\
&= \mathbf{1}^\top \operatorname{Diag}(\boldsymbol{b}_{[\boldsymbol{s}]})(A_{[\boldsymbol{s}]})^{-1} \operatorname{Diag}(\boldsymbol{b}_{[\boldsymbol{s}]})\mathbf{1} \\
&= \mathbf{1}^\top \left(\operatorname{Diag}(\mathbf{1}/\boldsymbol{b}_{[\boldsymbol{s}]}) A_{[\boldsymbol{s}]} \operatorname{Diag}(\mathbf{1}/\boldsymbol{b}_{[\boldsymbol{s}]})\right)^{-1} \mathbf{1} \\
&= \mathbf{1}^\top \left(\Sigma_{[\boldsymbol{s}]}\right)^{-1} \mathbf{1},
\end{aligned}$$

where $\Sigma = \operatorname{Diag}(\mathbf{1}/\boldsymbol{b}) A \operatorname{Diag}(\mathbf{1}/\boldsymbol{b})$. This implies that, under the minor assumption, solving the minimum-variance portfolio problem is equivalent to solving the core problem.

**Example 2. Column Subset Selection Problem:** Let $X \in \mathbb{R}^{m\times p}$ be a data matrix and define $X_{[:,\boldsymbol{s}]}$ as the submatrix with the selected columns associated with ones in $\boldsymbol{s}$. Further define the (orthogonal) projector

$$P_{\boldsymbol{s}} := X_{[:,\boldsymbol{s}]} \left(X_{[:,\boldsymbol{s}]}^\top X_{[:,\boldsymbol{s}]}\right)^\dagger X_{[:,\boldsymbol{s}]}^\top.$$

The *column subset selection problem (CSSP)* in Frobenius norm is

$$\min_{\boldsymbol{s}\in\{0,1\}^p:\,|\boldsymbol{s}|\le k} \|\, X - P_{\boldsymbol{s}} X \,\|_F^2. \tag{5}$$

Using $\|B\|_F^2 = \mathsf{Tr}(B^\top B)$ and the idempotence of $P_{\boldsymbol{s}}$, we obtain

$$\left\|X - P_{\boldsymbol{s}} X\right\|_F^2 = \|X\|_F^2 - \mathsf{Tr}\big(X^\top P_{\boldsymbol{s}} X\big),$$

which implies, (**??**) is equivalent to

$$\max_{\boldsymbol{s}\in\{0,1\}^p:\,|\boldsymbol{s}|\le k} \mathsf{Tr}\big(X^\top P_{\boldsymbol{s}} X\big). \tag{6}$$

Introduce a Rademacher vector $\boldsymbol{\xi} \in \{\pm 1\}^n$ with $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top] = I$ (i.e, generate elements of $\boldsymbol{\xi}$ independently and uniformly on $\{-1,+1\}$ ). By Hutchinson's identity **?**, $\mathsf{Tr}(B) = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}^\top B \boldsymbol{\xi}]$ for any square $B$, so (**??**) can be written as

$$\mathsf{Tr}\big(X^\top P_{\boldsymbol{s}} X\big) = \mathbb{E}_{\boldsymbol{\xi}}\Big[\boldsymbol{\xi}^\top X^\top P_{\boldsymbol{s}} X \boldsymbol{\xi}\Big]. \tag{7}$$

Now set

$$A := X^\top X \in \mathbb{R}^{n\times n}, \qquad \boldsymbol{b}^{(\boldsymbol{\xi})} := X^\top X\, \boldsymbol{\xi} \in \mathbb{R}^n,$$

so that $A_{[\boldsymbol{s}]} = X_{[:,\boldsymbol{s}]}^\top X_{[:,\boldsymbol{s}]}$ and $\boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})} = X_{[:,\boldsymbol{s}]}^\top X\xi$. Then the inner quadratic in (**??**) becomes

$$\boldsymbol{\xi}^\top X^\top P_{\boldsymbol{s}} X \boldsymbol{\xi} = \big(\boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})}\big)^\top \big(A_{[\boldsymbol{s}]}\big)^\dagger \boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})},$$

and the CSSP objective can be expressed as the following expected instance of our core problem:

$$\max_{\boldsymbol{s}\in\{0,1\}^p:\,|\boldsymbol{s}|\le k} \mathbb{E}_{\boldsymbol{\xi}}\Big[\, \big(\boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})}\big)^\top \big(A_{[\boldsymbol{s}]}\big)^\dagger \boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})} \,\Big]. \tag{8}$$

Daniel will work on column subset selection problem focusing on:

- Rewrite the following results,
- Rewrite the algorithm,
- Investigate improvements techniques,
- Running simulations to compare with existing methods.

# 2 CSSP - Daniel

## 2.1 Boolean Relaxation

We make the following key assumption.

**Assumption 1:** *Matrix $A = X^\top X$ is positive definite.*

This assumption is not restricitive, because we can perform a ridge regularization, by replacing $A$ with $A + \lambda I$ for some $\lambda > 0$ (Moka et al., 2025). ( Also cited (Fastrich, Paterlini and Winker, 2015))

Now, we show that all elements of $\boldsymbol{b}^{(\boldsymbol{\xi})} = X^\top X \boldsymbol{\xi} = A\boldsymbol{\xi}$ is non-zero.

$$\boldsymbol{b}^{(\boldsymbol{\xi})} = X^\top X \boldsymbol{\xi} = A\boldsymbol{\xi} \tag{9}$$

$$\iff A^{-1\boldsymbol{\xi}_{-}} \tag{10}$$

Note that, since $A$ is postive definite, Since $\boldsymbol{\xi} \in \{\pm 1\}^n$, all elements of $\boldsymbol{b}^{(\boldsymbol{\xi})} = X^\top X \boldsymbol{\xi}$ is non-zero.

This is because, since $A = X^\top X \succ 0$

We can reformulate (**??**) into

$$\max_{\boldsymbol{s} \in \{0,1\}^p :\, |\boldsymbol{s}| \le k} \mathbb{E}_{\boldsymbol{\xi}} \Big[ \mathbf{1}^\top (\Sigma_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})})^{-1} \mathbf{1} \Big], \tag{11}$$

where

$$\Sigma_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})} := \mathrm{Diag}\,(\mathbf{1}/\boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})}) \big(A_{[\boldsymbol{s}]}\big)\, \mathrm{Diag}\,(\mathbf{1}/\boldsymbol{b}_{[\boldsymbol{s}]}^{(\boldsymbol{\xi})})$$

assuming that $A$ is a full rank, using the ”crucial obsservation” from earlier.

*We might not be able to make the full-rank assumption, and also have to prove that b is all non-zero*

*Also, clearly state what $f_\delta(t)$ is.*

Now we expect to prove that Boolean relaxation of (**??**) is is indeed a boolean relaxation of (**??**) thanks to the fact that expectation operator is a linear operator.

# 3 Boolean relaxation

We now provide a Boolean relaxation of (**??**) as an auxiliary continuous function on $[0,1]^p$, controlled by a tuning parameter $\delta > 0$. [This stuff is from Moka et al (2025) on portfolio optimization]. To simplify the notation, define

$$T_{\boldsymbol{t}} = \mathrm{Diag}\,(\boldsymbol{t}) \quad \text{and} \quad \widetilde{\Sigma}_{\boldsymbol{t}} = T_{\boldsymbol{t}} \Sigma T_{\boldsymbol{t}} + \delta(I - T_{\boldsymbol{t}}^2). \tag{12}$$

Then, our proposed Boolean relaxation of (**??**) is given by

$$\min_{\boldsymbol{t} \in \mathcal{C}_k} f_\delta(\boldsymbol{t}), \quad \text{where} \quad f_\delta(\boldsymbol{t}) = -\boldsymbol{t}^\top \widetilde{\Sigma}_{\boldsymbol{t}}^{-1} \boldsymbol{t}, \tag{13}$$

and for each $k$ the constraint set $\mathcal{C}_k$ is a polytope defined as

$$\mathcal{C}_k = \{\boldsymbol{t} \in [0,1]^p : \boldsymbol{t}^\top \mathbf{1} \le k\}. \tag{14}$$

The following result, Theorem **??**, shows why (**??**) is a relaxation of the target problem (**??**). It shows that $f_\delta(\boldsymbol{t})$ is continuous on the hypercube $[0,1]^p$ and its shape can be controlled by the auxiliary parameter $\delta$ while keeping the values of $f_\delta(\boldsymbol{t})$ fixed—independent of $\delta$—at all the (binary) corners $\boldsymbol{s} \in \{0,1\}^p$. In addition, (iii) shows that $f_\delta(\boldsymbol{t})$ increases with $\delta$ for any fixed interior point $\boldsymbol{t}$, while (iv) shows that the optimum of (**??**) is on a simplex.

**Theorem 1** (Theorems 2 & 3 of **?**)**.** *The following hold:*

*(i) The objective function $f_\delta(\boldsymbol{t})$ in* (**??**) *is continuous on $[0,1]^p$.*

*(ii) For every binary vector $\boldsymbol{s} \in \{0,1\}^p$ (i.e., a corner point on the hypercube $[0,1]^p$),*

$$f_\delta(\boldsymbol{s}) = -\mathbf{1}^\top \Sigma_{[\boldsymbol{s}]}^{-1} \mathbf{1}, \quad \textit{for all } \delta > 0.$$

*(iii) For every fixed $\boldsymbol{t} \in (0,1)^p$, $f_\delta(\boldsymbol{t})$ is monotonically increasing in $\delta > 0$.*

*(iv) For any $k = 1, \ldots, p$ and $\delta > 0$,*

$$\min_{\boldsymbol{t} \in \mathcal{C}_k} f_\delta(\boldsymbol{t}) = \min_{\boldsymbol{t} \in \mathcal{S}_k} f_\delta(\boldsymbol{t}),$$

*here, the simplex $\mathcal{S}_k = \{\boldsymbol{t} \in [0,1]^p : \boldsymbol{t}^\top \mathbf{1} = k\}$ corresponds to the polytope $\mathcal{C}_k$ given in* (**??**)*.*

*(v) Let $\eta_1$ be the largest eigenvalue of $\Sigma$. Then, $f_\delta(\boldsymbol{t})$ strictly concave over $[0,1]^p$ for $\delta \geq \eta_1$.*

### 3.1 Gradient

We first derive a convenient closed–form expression for the gradient of $f_\delta$.

**Lemma 1** (Gradient of the relaxed objective)**.** *For each $\boldsymbol{t} \in [0,1]^p$ and define $\boldsymbol{x} := \widetilde{\Sigma}_{\boldsymbol{t}}^{-1} \boldsymbol{t}$ and $\boldsymbol{z} := \Sigma(\boldsymbol{t} \odot \boldsymbol{x})$. Then $f_\delta$ is differentiable at $\boldsymbol{t}$ and*

$$\nabla f_\delta(\boldsymbol{t}) = -2\,\boldsymbol{x} \;+\; 2\,\boldsymbol{x} \odot \boldsymbol{z} \;-\; 2\delta\,\boldsymbol{t} \odot \boldsymbol{x} \odot \boldsymbol{x}. \tag{15}$$

*Moreover, with $\Pi_{\boldsymbol{t}} := \Sigma + \delta(T_{\boldsymbol{t}}^{-2} - I)$, on the interior points, the gradient admits the equivalent form*

$$\nabla f_\delta(\boldsymbol{t}) = -2\delta\, \frac{(\Pi_{\boldsymbol{t}}^{-1}\mathbf{1})^2}{\boldsymbol{t}^3}, \qquad \boldsymbol{t} \in (0,1)^p. \tag{16}$$

*where all operations between vectors are elementwise.*

## 4 Algorithms for KKT-Minimal Points

In this section, we develop several continuous optimization algorithms for the relaxed problem (**??**) over the simplex $\mathcal{S}_k$. The algorithms operate by updating the regularization parameter $\delta$ during the iterations and are designed to converge to KKT binary corner of $\mathcal{S}_k$ at $\delta = \eta_1$, yielding a candidate solution for the original sparse portfolio problem.

### 4.1 Frank-Wolfe Homotomy Method

Algorithm **??** is a variant of the standard Frank-Wolfe algorithm, similar to the `Grid-FW` of **?**. This algorithm is coupled with a continuation scheme in the regularization parameter $\delta$.

[Explain the algorithm.]

**Algorithm 1** `FW-Homotopy`$(\Sigma, k, \alpha, n)$

```
Compute the largest and smallest eigenvalues $\eta_1$ and $\eta_p$ of $\Sigma$
$\tau \leftarrow 10^{-4}$                                  ▷ Tolerance for termination
$\varepsilon \leftarrow 0.1(k/p)$
$\delta_0 \leftarrow 3\eta_p\varepsilon^2/(1+3\varepsilon^2)$
$r \leftarrow (\eta_1/\delta_0)^{1/(n-1)}$
$\boldsymbol{t} \leftarrow (k/p)\mathbf{1}$
$\ell \leftarrow 1$
repeat
    $\delta \leftarrow \delta_0 r^\ell$
    Compute the gradient $\nabla f_\delta(\boldsymbol{t})$
    Let $\boldsymbol{s} \in \{0,1\}^p$ have ones at the positions of the $k$ smallest components of $\nabla f_\delta(\boldsymbol{t})$
    $\boldsymbol{t} \leftarrow (1-\alpha)\boldsymbol{t} + \alpha\boldsymbol{s}$
    $\ell \leftarrow \ell + 1$
    if $\min_{j=1,\dots,p} \min\{t_j,\, 1-t_j\} \le \tau$ then
        Set $\delta \leftarrow \eta_1$ and compute $\boldsymbol{g} \leftarrow \nabla f_\delta(\boldsymbol{s})$
        if $\max_{j:\, s_j=1} g_j \le \min_{i:\, s_i=0} g_i$ then          ▷ KKT-Certification at $\delta = \eta_1$
            return $\boldsymbol{s}$
        end if
    end if
until $\ell > n$
return $\boldsymbol{s}$
```

# A Proofs

*Proof of Lemma* **??**. For $\boldsymbol{t} \in [0,1]^p$, since $\widetilde{\Sigma}_{\boldsymbol{t}}$ invertible, using the identity $\partial A^{-1} = -A^{-1}(\partial A)A^{-1}$ for matrix differentials and the product rule, one obtains

$$\frac{\partial f_\delta}{\partial t_i} = -2x_i \; + \; 2x_i z_i \; - \; 2\delta\, t_i x_i^2,$$

where $\boldsymbol{x} = \widetilde{\Sigma}_{\boldsymbol{t}}^{-1}\boldsymbol{t}$ and $\boldsymbol{z} = \Sigma(\boldsymbol{t} \odot \boldsymbol{x})$. Collecting the components yields (**??**), which is well-defined for all such $\boldsymbol{t} \in [0,1]^p$ (no division by $t_i$ is involved).

For the second expression, restrict to interior points $\boldsymbol{t} \in (0,1)^p$, so that $T_{\boldsymbol{t}}$ is invertible. Then

$$\widetilde{\Sigma}_{\boldsymbol{t}} = T_{\boldsymbol{t}}\Pi_{\boldsymbol{t}}T_{\boldsymbol{t}}, \qquad \text{with } \Pi_{\boldsymbol{t}} := \Sigma + \delta(T_{\boldsymbol{t}}^{-2} - I),$$

and hence

$$\widetilde{\Sigma}_{\boldsymbol{t}}^{-1}\boldsymbol{t} = T_{\boldsymbol{t}}^{-1}\Pi_{\boldsymbol{t}}^{-1}T_{\boldsymbol{t}}^{-1}\boldsymbol{t} = T_{\boldsymbol{t}}^{-1}\Pi_{\boldsymbol{t}}^{-1}\mathbf{1}.$$

Thus

$$x_i = \frac{\left(\Pi_{\boldsymbol{t}}^{-1}\mathbf{1}\right)_i}{t_i}, \qquad t_i > 0.$$

Substituting this representation of $x_i$ (and the corresponding expression for $z_i$) into (**??**), the terms $-2x_i + 2x_i z_i$ cancel, yielding

$$\frac{\partial f_\delta}{\partial t_i} = -2\delta\, t_i x_i^2 = -2\delta\, \frac{\left(\Pi_{\boldsymbol{t}}^{-1}\mathbf{1}\right)_i^2}{t_i^3}, \quad i = 1, \dots, p.$$

Writing this in vector form gives the stated alternative gradient expression for $\boldsymbol{t} \in (0,1)^p$. □