Educational Testing Problem and Non-smooth Optimization

Moody T. Chu
and
Joel W. Wright

Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205

Outline

- Statistics Background
- Ellipsoid Method
    - ◇ Reformulation
    - ◇ Subgradient
- Dikin's Method
    - ◇ Barrier Function
    - ◇ Inner Ellipsoid
    - ◇ Analytic Center
- Numerical Experiment

What is the Problem?

- Educational test problem:

$$\begin{aligned} \text{Maximize} \quad & \mathrm{trace}(D), & (1) \\ \text{Subject to} \quad & S - D \geq 0,\ D \geq 0 & (2) \end{aligned}$$

  ◇ S; symmetric, positive definite (≥ 0).

  ◇ $D = \mathrm{Diag}\{d_1, \ldots, d_n\}$; diagonal.

- Convex programming problem:

$$\begin{aligned} \text{Minimize} \quad & \phi(x) & (3) \\ \text{subject to} \quad & \psi(x) \leq 0. & (4) \end{aligned}$$

  ◇ $\psi = [\psi_1, \ldots, \psi_m]^T$.

  ◇ ϕ, $\psi_1, \ldots, \psi_m$; convex.

Basic Terminology in Statistics

- X; scalar random variable.
- $f(x)$; probability distribution function of X.
- $\mu_X := \mathcal{E}(X) := \Sigma_x\, x f(x)$; mean value of X.
- $\sigma_X^2 := \mathcal{E}((X-\mu_X)^2)$; variance of X.
- $\text{cov}(X,Y) := \mathcal{E}((X-\mu_X)(Y-\mu_Y))$; covariance of X and Y.
- $\rho := \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$; correlation coefficient.
  - ⋄ X, Y stochastically independent $\Longrightarrow \rho = 0$.

Ideal case description

- $\hat{X}$; random variables of total score in one course.
- $\hat{T}$; random variables of "true" total score, (hypothetical, unknown).
- $\hat{E} = \hat{X} - \hat{T}$.
- Assume $\mathcal{E}(\hat{E}) = 0$, (So, $\mu_{\hat{X}} = \mu_{\hat{T}}$).
- $\hat{X}_1, \ldots, \hat{X}_N$; a class of N students (Observed scores).
  - ◇ $\overline{\hat{X}} := \frac{\Sigma_i \hat{X}_i}{N}$; sample mean, $(= \mu_{\hat{T}})$.
  - ◇ $s_{\hat{X}} := \Sigma_i \frac{(\hat{X}_i - \mu_{\hat{T}})^2}{N}$; sample variance.
- $\hat{X}^{(1)}, \hat{X}^{(2)}$; random variable of total score in two identical courses.
  - ◇ $\rho := \frac{1}{N} \frac{\Sigma_i (\hat{X}_i^{(1)} - \mu_{\hat{T}})(\hat{X}_i^{(2)} - \mu_{\hat{T}})}{s_{\hat{X}^{(1)}} s_{\hat{X}^{(2)}}}$; reliability coefficient.
  - ◇ $\rho \approx 1 \Longrightarrow$ observed total scores are very reliable.
  - ◇ Under reasonable assumption,

$$\rho \approx \frac{\sigma_{\hat{T}}^2}{\sigma_{\hat{X}}^2}.$$

Practical case description

- To receive information about $\hat{X}^{(1)}$ and $\hat{X}^{(2)}$,
  - ◇ The identical course must be run twice.
  - ◇ Impossible unless introducing complete amnesia in the students after each test.
- Desire to deduce information from one single course.
  - ◇ **The course itself contains m tests.**
- New notation:
  - ◇ X_j; random variable on score of test j.
  - ◇ x_{ij}; the score of student i on test j.
  - ◇ $\Omega_X := [\mathrm{cov}(X_j, X_k)]$; covariance matrix.
    - ▷ symmetric, positive definite.
  - ◇ Similar definition for hypothetical true scores T_j and errors E_j.

- Assume stochastic independence between
  - ◇ E_j and T_k for all j, k,
  - ◇ E_j and E_k for all $j \neq k$.

  Then
$$\Omega_X = \Omega_T + \Omega_E$$

- Observe
  - ◇ $\sigma^2_{\hat{T}} = \Sigma_{ij} \operatorname{cov}(T_i, T_j)$ (since $\hat{T} = \sigma_j T_j$).
  - ◇ Similarly for $\sigma^2_{\hat{X}}$ and $\sigma^2_{\hat{E}}$.
  - ◇ Ω_E is diagonal. ($\sigma^2_{\hat{E}} = \operatorname{trace}(\Omega_E)$.)
  - ◇ $\rho = 1 - \frac{\sigma^2_{\hat{E}}}{\sigma^2_{\hat{X}}}$.
- **Maximize $\phi =$ trace(Ω_E) in $1 - \frac{\phi}{\sigma^2_{\hat{X}}}$ gives a lower bound on ρ.**
- Don't forget the condition

$$\begin{aligned} \Omega_{\hat{T}} = \Omega_{\hat{X}} - \Omega_{\hat{E}} &\geq 0, \\ \Omega_{\hat{E}} &\geq 0. \end{aligned}$$

Fletcher's Idea (85)

- Study the geometric structure of the feasible domain, including expressions for the normal cone, feasible directions and optimality conditions.
- Replace the constraint (2) by a set of nonlinear algebraic equations.
- SQP defined by a guess of the nullity of $S - D$ and a penality function.
- Very difficult to program.

Reformulation

- Fix D,
  - $\diamond$ $\lambda S - D \geq 0 \iff \lambda I - S^{-1/2} D S^{-1/2} \geq 0$.
  - $\diamond$ $S - D \geq 0 \iff \lambda_1(D - S) \leq 0$.
    - $\triangleright$ $\lambda_1 =$ largest eigenvalue.
- Define:

$$
\begin{aligned}
\mu(D) &:= \operatorname{argmin}_\lambda \{\lambda S - D \geq 0\} \\
&= \lambda_1(S^{-1/2} D S^{-1/2}), \\
\nu(D) &:= \lambda_1(D - S).
\end{aligned}
$$

- Observe:
  - $\diamond$ μ, ν; convex.
  - $\diamond$ $\frac{D}{\mu(D)}$; diagonal, invariant under scalar multiplication.

- First reformulation:

$$\begin{aligned} \text{Minimize} \quad & \mu(D), && (5) \\ \text{subject to} \quad & \operatorname{trace}(D) = 1,\ -D \le 0. && (6) \end{aligned}$$

- Second reformulation:

$$\begin{aligned} \text{Minimize} \quad & -\operatorname{trace}(D), && (7) \\ \text{subject to} \quad & \nu(D) \le 0,\ -D \le 0. && (8) \end{aligned}$$

- Getting inequality constraints:

$$\begin{aligned} \text{Minimize} \quad & \tilde{\mu}(d_1, \ldots, d_{n-1}), && (9) \\ \text{subject to} \quad & -d_i \le 0,\ \sum_{i=1}^{n-1} d_i - 1 \le 0 && (10) \end{aligned}$$

  ◇ $\tilde{\mu}(d_1, \ldots, d_{n-1}) := \mu\left(d_1, \ldots, d_{n-1}, 1 - \Sigma_{i=1}^{n-1} d_i\right)$.

- Trouble!

  ◇ Eigenvalue optimization $\Longrightarrow$ Non-differentiable at coalescent points.

  ◇ Quite so often at an optimal solution the eigenvalues coalesce.

Subgradient

- $f : R^n \longrightarrow R$; convex,
  - ◇ $g \in R^n$, a subgradient if
$$f(z) \geq f(x) + g^T(z - x) \text{ for all } z.$$
  - ◇ Subgradient always exists.
- Important implication:
$$f(z) > f(x) \text{ whenever } g^T z > g^T x.$$
- The "usable" half-space:
$$H(x, g) := \left\{ z \in R^n | g^T(z - x) \leq 0 \right\}.$$

Computing Subgradient

- THEOREM (OVERTON (92)): Let $q(D)$ be any normalized eigenvector of $\lambda_1(A(D))$. Then the vector $g = [g_1, \ldots, g_n]^T$ with

$$g_k := q(D)^T A_k(D) q(D)$$

is a subgradient of $\lambda_1(A(D))$.

 ◇ $A_k(D) := \frac{\partial A(D)}{\partial d_k}$.

- In our case,

$$\begin{aligned} \partial\mu(D) &= \left(S^{-1/2}q(D)\right) \circ \left(S^{-1/2}q(D)\right), \\ \partial\nu(D) &= q(D) \circ q(D), \\ \partial\tilde{\mu}(D) &= T\partial\mu(D) \end{aligned}$$

 ◇ $\circ$; Hadamard product.
 ◇

$$T := \begin{bmatrix} 1 & 0 & \ldots & 0 & -1 \\ 0 & 1 & & & -1 \\ \vdots & & \ddots & & \vdots \\ 0 & & & 1 & -1 \end{bmatrix}.$$

Ellipsoid Method

- Characterization of an ellipsoid,

  $$E = E(B, a) := \left\{ x \in R^n | (x-a)^T B^{-1} (x-a) \leq 1 \right\}.$$

  ◇ $a \in R^n$; the center.

  ◇ B; symmetric, positive definite.

- The method (Shor (77), Khachiyan (80)):

  ◇ $E^{(1)}$ contains the feasible minimizer x^*.

  ◇ $E^{(k+1)}$ is the ellipsoid of minimum volume (the Löwner-John ellipsoid) that contains the half-sliced ellipsoid $E^{(k)} \cap H(x^{(k)}, g^{(k)})$ where

  $$g^{(k)} := \begin{cases} \partial\phi(x^{(k)}), & \text{if } \psi(x^{(k)}) \leq 0; \\ \partial\psi(x^{(k)}), & \text{if } \psi(x^{(k)}) > 0. \end{cases} \tag{11}$$

What is the Idea?

- If $x^{(k)}$ is feasible, then we discard all points where objective values are greater than or equal to $\phi(x^{(k)})$.
- If $x^{(k)}$ is not feasible, then we discard all points which are further guaranteed to be infeasible.
- $B^{(k+1)}$ and $x^{(k+1)}$ can be explicitly described in terms of $B^{(k)}$, $x^{(k)}$ and $g^{(k)}$.

The Algorithm (Bland, Goldfarb & Todd, (81))

Given $B^{(1)}$ and $x^{(1)}$ so that $E^{(1)}$ contains a feasible minimizer, do:

Compute $\psi(x^{(k)})$,

If $\psi(x^{(k)}) > 0$,

Compute any $g^{(k)} \in \partial\psi(x^{(k)})$;

$\gamma := \sqrt{{g^{(k)}}^T B^{(k)} g^{(k)}}$;

$g := \frac{g^{(k)}}{\gamma}$;

If $\psi(x^{(k)}) - \gamma > 0$, quit.

Else,

Compute any $g^{(k)} \in \partial\phi(x^{(k)})$;

$\gamma := \sqrt{{g^{(k)}}^T B^{(k)} g^{(k)}}$;

$g := \frac{g^{(k)}}{\gamma}$;

$b := B^{(k)} g$;

$x^{(k+1)} := x^{(k)} - \frac{b}{n+1}$;

$B^{(k+1)} := \frac{n^2}{n^2-1}\left(B^{(k)} - \frac{2}{n+1} b b^T\right)$;

If $\psi(x^{(k)}) \le 0$ and $\gamma \le \epsilon$, stop.

Interesting Features

- Global convergence,

$$\frac{\mathrm{vol}(E^{(k+1)})}{\mathrm{vol}(E^{(k)})} = \left(\left(\frac{n}{n+1}\right)^{n+1}\left(\frac{n}{n-1}\right)^{n-1}\right)^{\frac{1}{2}} < e^{\frac{-1}{2n}}.$$

- Finite precision arithmetic $\Longrightarrow B^{(k)}$ indefinite.
- Unstability can be remedied by updating the factor of $B^{(k)} = J^{(k)}J^{(k)^T}$.
  - ◇ Square root is avoided.
  - ◇ Algorithm can easily be modified.
- Cholesky factor $J^{(k)}$ can be updated at $O(n^2)$ flops.

Barrier Function

- Define
$$\phi(D) := \sum_{i=1}^{n} \ln \frac{1}{\lambda_i(D)} + \sum_{i=1}^{n} \ln \frac{1}{d_i}.$$
  - ◇ $\lambda_i(D) =$ the i^{th} eigenvalue of $S - D$.
  - ◇ Defined only for strictly feasible D.
  - ◇ The level curves of ϕ should be reasonable approximations to the boundary of the feasible domain.
- Derivatives of ϕ,
$$\begin{aligned} \phi(D) &= \ln\det(S-D)^{-1} + \ln\det D^{-1}, \\ \nabla\phi(D) &= \operatorname{diag}\left((S-D)^{-1} - D^{-1}\right), \\ H(D) &= (S-D)^{-1} \circ (S-D)^{-1} + D^{-1} \circ D^{-1}. \end{aligned}$$
- Schur product theorem $\Longrightarrow H(D) \geq 0 \Longrightarrow \phi$ strictly convex.

Inner Ellipsoid

- THEOREM: Suppose $D^{(c)}$ is a strictly feasible point with respect to (2). Then every diagonal matrix $D^{(+)}$ with $d^{(+)}$ from the ellipsoid $E(H(D^{(c)})^{-1}, d^{(c)})$ is also feasible.

Dikin's Method

- Suppose $D^{(c)}$ is feasible, approximate the (ETP) **locally** by

$$\begin{aligned} \text{Maximize} \quad & e^T d, && (12)\\ \text{subject to} \quad & d \in E(H(D^{(c)})^{-1}, d^{(c)}). && (13) \end{aligned}$$

  ◇ $d^{(c)} = \operatorname{diag}(D^{(c)})$.

  ◇ $e := [1, \ldots, 1]^T$.

- LEMMA: For $p \neq 0$, the maximal value of $p^T x$ subject to the condition $x \in E(B, a)$ occurs at

$$x^* := a + \frac{1}{\sqrt{p^T B p}} B p.$$

- The algorithm:

  Given $x^{(1)} \in R^n$ strictly feasible, do:

  If $A - \operatorname{Diag}(x^{(k)})$ is singular, stop

  Else,

  Solve $H(\operatorname{Diag}(x^{(k)}))b = e$ for b;

  $x^{(k+1)} := x^{(k)} + \frac{1}{\sqrt{e^T b}} b$.

Analytic Center

- $\hat{D} := \mathrm{argmin}_D \phi(D) =$ analytic center.
- $\nabla\phi(\hat{D}) = \mathrm{diag}(S - \hat{D})^{-1} - \hat{D}^{-1} = 0.$
- Compute the center by a Newton method with damping (Nesterov and Nemirovsky's Method):

  Given any initial point $D^{(0)}$ that is feasible, do:

  Compute $\nabla\phi(D^{(k)})$ and $H(D^{(k)})$;

  Solve $H(D^{(k)})^{\frac{1}{2}}\xi = \nabla\phi(D^{(k)})$;

  $\rho := \|\xi\|$,

  If $\rho \leq \frac{1}{4}$,

  $\alpha^{(k)} := 1$;

  Else,

  $\alpha^{(k)} := \frac{1}{1+\rho}$.

  Solve $H(D^{(k)})\delta^{(k)} = -\alpha^{(k)}\nabla\phi(D^{(k)})$;

  $D^{(k+1)} := D^{(k)} + \mathrm{Diag}(\delta^{(k)})$.
- Damping factor $\alpha^{(k)}$ results in $D^{(k+1)}$ being feasible.

Numerical Experiment

- Test data:
  - $\diamond$ $X = [x_{ij}] \in R^{N \times m}$.
  - $\diamond$ x_{ij} = the score of student i on test j.
  - $\diamond$ $v = \{v_1, \ldots, v_n\}$ = subset of column indices being considered. $(n \leq m)$.
  - $\diamond$ $S = S_v = [s_{ij}]$

$$s_{jk} := \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{iv_j} - \overline{x}_{v_j}\right) \left(x_{iv_k} - \overline{x}_{v_k}\right)$$

  - $\diamond$ $\overline{x}_j = \Sigma_{i=1}^{N} x_{ij}/N$ = mean of test j.
- Computing the analytic center:
  - $\diamond$ Starting point $D^{(0)} := 0.9\lambda_n(S)I$.
  - $\diamond$ Stopping criterion $\leq \epsilon\|S\|$.