MIT 15.095 M. L. under a Modern Optimization Lens - Cheat Sheet

Author : David Alexandre Nze Ndong

1. The Optimization Lenses

Mixed Integer Optimization :

$$\begin{aligned} &\max \mathbf{c}^\top \mathbf{x} + \mathbf{h}^\top \mathbf{y} \\ &s.t.\ \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \le \mathbf{b} \\ &\quad \mathbf{x} \in \mathbb{Z}_+^n \\ &\quad \mathbf{y} \in \mathbb{R}_+^m \end{aligned}$$

Mixed Integer Quadratic Optimization :

$$\begin{aligned} &\max \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{h}^\top \mathbf{y} \\ &s.t.\ \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} \le \mathbf{b} \\ &\quad \mathbf{x} \in \mathbb{Z}_+^n \\ &\quad \mathbf{y} \in \mathbb{R}_+^m \end{aligned}$$

Convex Optimization :

$$\begin{aligned} &\min f(\mathbf{x}) \\ &s.t.\ g_i(\mathbf{x}) \le 0,\ \forall i \in [m] \end{aligned}$$

Semi-definite Optimization :

$$\begin{aligned} &\min \langle \mathbf{C}, \mathbf{X} \rangle \\ &s.t.\ \langle \mathbf{A_i}, \mathbf{X} \rangle = \mathrm{Tr}(\mathbf{C}^\top \mathbf{X}) \ge b_i,\ \forall i \in [m] \\ &\quad \mathbf{X} \succcurlyeq \mathbf{0}, \end{aligned}$$

Second Order Cone Problem:

$$\begin{aligned} &\min \mathbf{c}^\top \mathbf{x} \\ &s.t.\ \|\mathbf{A_i}\mathbf{x} + \mathbf{b_i}\|_2 \le \mathbf{f_i}^\top \mathbf{x} + d_i \quad \forall i \in [m] \\ &\quad \mathbf{x} \ge \mathbf{0}, \end{aligned}$$

Robust Optimization:

$$\begin{aligned} &\min \mathbf{c}^\top \mathbf{x} \\ &s.t.\ \mathbf{a_i}^\top \mathbf{x} \ge b_i, \quad \forall \mathbf{a}_i \in \mathcal{U}_i,\ \forall i \in [m] \\ &\quad \mathbf{x} \ge \mathbf{0}, \end{aligned}$$

What is theoretical tractability: The problem can be solved in *polynomial time* in the bits $\mathcal{P}$, in contrast to $\mathcal{NP}$-hard problem. **What is practical tractability:** The problem can be solved in *time and for size* that are suitable with the problem

2. Robust Regression

Formulation :

$$\min_{\boldsymbol{\beta}} \max_{\substack{\boldsymbol{\delta} \in \mathcal{V} \\ \boldsymbol{\Delta} \in \mathcal{U}}} g(\mathbf{y} + \boldsymbol{\delta} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})$$

$$\equiv$$

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}} \bar{g}(\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})$$

Norms and Dual norms :
dual-norm $\|\cdot\|_* = \max_{\mathbf{x} \in \mathbb{R}^n} \frac{\boldsymbol{\beta}^\top \mathbf{x}}{\|\mathbf{x}\|} \mid \max_{\mathbf{A} \in \mathbb{R}^{n \times p}} \frac{\langle \mathbf{A}, \boldsymbol{\Delta} \rangle}{\|\mathbf{A}\|}$

ℓ_r*-norm* or r-*Frobenius norm* $\|\cdot\|_r = \left(\sum_{i \in [n]} |\beta_i|^r\right)$

p-spectral (Schatten) norm $\|\boldsymbol{\Delta}\|_{\sigma_r} = \|\boldsymbol{\mu}(\boldsymbol{\Delta})\|_r$
(with $\boldsymbol{\mu}(\boldsymbol{\Delta})$ the vector of singular values of $\boldsymbol{\Delta}$)

induced-norm $\|\cdot\|_{(h,g)} = \max_{\mathbf{x} \in \mathbb{R}^n} \frac{g(\boldsymbol{\Delta}\boldsymbol{\beta})}{h(\boldsymbol{\beta})}$

Thm: if $g : \mathbb{R}^n \mapsto \mathbb{R}$ is a seminorm, which is not identically zero and $h : \mathbb{R}^n \mapsto \mathbb{R}$ is a norm, then for any $\mathbf{z} \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\max_{\boldsymbol{\Delta} \in \mathcal{U}_{(h,g;\lambda)}} g(\mathbf{z} + \boldsymbol{\Delta}\boldsymbol{\beta}) = g(\mathbf{z}) + \lambda h(\boldsymbol{\beta})$$

Cor:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{(q,r;\lambda)}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})\|_r = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_r + \lambda \|\boldsymbol{\beta}\|_q$$

Thm:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \mathcal{U}_{Fr;\lambda}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta})\|_r = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_r + \lambda \|\boldsymbol{\beta}\|_{r^*}$$

(Can be extended to `Elastic Net`)

3. Sparse Regression

Formulation:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}), \quad \text{s.t. } \|\beta\|_0 \le k$$

Primal Approach : First Order Algorithm

Noticing that g is convex ($\boldsymbol{\nabla}^2 g(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{X} + \mathbf{I}/\gamma \succcurlyeq 0$) with *Lipschitz continuous gradient*:

$$\|\boldsymbol{\nabla} g(\boldsymbol{\beta}) - \boldsymbol{\nabla} g(\tilde{\boldsymbol{\beta}})\| \le \underbrace{\lambda_{\max}(\mathbf{X}^\top \mathbf{X} + \mathbf{I}/\gamma)}_{\ell} \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|$$

We have the following upper bound ($\forall L \ge \ell$)

$$g(\boldsymbol{\eta}) \le Q_L(\boldsymbol{\eta}, \boldsymbol{\beta}) \mathrel{\hat{=}} g(\boldsymbol{\beta}) + \frac{L}{2}\|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2 + \boldsymbol{\nabla} g(\boldsymbol{\beta})^\top (\boldsymbol{\eta} - \boldsymbol{\beta})$$

We just need to compute (*Disc. First Order Iter.*):

$$\begin{aligned} \boldsymbol{\eta}(\boldsymbol{\beta}_m) &= \operatorname{argmin}_{\|\boldsymbol{\eta}\|_0 \le k} Q_L(\boldsymbol{\eta}, \boldsymbol{\beta}_m) \\ &= \operatorname{argmin}_{\|\boldsymbol{\eta}\|_0 \le k} \left\| \boldsymbol{\eta} - \left(\boldsymbol{\beta}_m - \frac{1}{L} \boldsymbol{\nabla} g(\boldsymbol{\beta}_m) \right) \right\|_2^2 \\ &= \mathbf{H}_k \left(\boldsymbol{\beta}_m - \frac{1}{L} \boldsymbol{\nabla} g(\boldsymbol{\beta}_m) \right) \mathrel{\hat{=}} \boldsymbol{\beta}_{m+1} \end{aligned}$$

Cvg. : $\min_{N \in [N]} \|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_{m+1}\|_2^2 \le \frac{2(g(\boldsymbol{\beta}_1) - g(\boldsymbol{\beta}^*))}{N(L - \ell)}$

Dual Approach :

$$\min \frac{1}{2}\mathbf{y}^\top \left(\mathbf{I_n} + \gamma \sum_{\mathbf{j} \in [\mathbf{p}]} \mathbf{s_j} \mathbf{X_j} \mathbf{X_j}^\top \right)^{-1} \mathbf{y}$$

We obtain a CIO which is computationally expensive and not many solvers are available currently. We usually solve this MIO approximative problems using a cutting plane method (we compute c the dual objective function).

Outer Approximation Algorithm

$$\begin{matrix} \mathbf{s}_{t+1} \\ \eta_{t+1} \end{matrix} = \left\{ \begin{matrix} \operatorname{argmin}_{s,\eta} & \eta \\ \text{s.t.} & \eta \ge c(\mathbf{s}_i) + \boldsymbol{\nabla} c(\mathbf{s}_i)^\top (\mathbf{s} - \mathbf{s}_i) \\ & \forall i \in [t], \mathbf{s} \in S_k^p \end{matrix} \right\}$$

4. Nonlinear Regression

1) *Convex Regression:* $\min_{f\in\mathcal{C}} \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$

$$f(\mathbf{x}_i) + \boldsymbol{\xi}_i^\top(\mathbf{x}_j - \mathbf{x}_i) \leq f(\mathbf{x}_j) \quad \text{(Conx. Prop.)}$$

$$f(\mathbf{x}) = \max_{i\in[n]}(f(\mathbf{x}_i) + \boldsymbol{\xi}_i^\top(\mathbf{x} - \mathbf{x}_i)) \quad \text{(Func. Hyp.)}$$

Formulation:

$$\begin{aligned} \min_{\boldsymbol{\theta},\{\boldsymbol{\xi}_i\}_{i\in[n]}} \ & \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2 + \frac{1}{2\gamma}\sum_{i=1}^{n}\|\boldsymbol{\xi}_i\|^2 \\ \text{s.t.}\ & \theta_i + \boldsymbol{\xi}_i^\top(\mathbf{x}_j - \mathbf{x}_i) \leq \theta_j, \quad \forall i,j\in[n] \\ & \boldsymbol{\theta}\in\mathbb{R}^n,\ \boldsymbol{\xi}_i\in\mathbb{R}^p \quad \forall i\in[n] \end{aligned}$$

How to solve: *Cutting plane algorithm, Delayed constraint generation*

2) *Sparse Convex Regression:*

Formulation:

$$\begin{aligned} \min_{\boldsymbol{\theta},\{\boldsymbol{\xi}_i\}_{i\in[n]}} \ & \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2 + \frac{1}{2\gamma}\sum_{i=1}^{n}\|\boldsymbol{\xi}_i\|^2 \\ \text{s.t.}\ & \theta_i + \boldsymbol{\xi}_i^\top(\mathbf{x}_j - \mathbf{x}_i) \leq \theta_j, \quad \forall i,j\in[n] \\ & \text{Supp}(\boldsymbol{\xi}_i)\subseteq S, \quad \forall i\in[n], \\ & \boldsymbol{\theta}\in\mathbb{R}^n,\ \boldsymbol{\xi}_i\in\mathbb{R}^p, \forall i\in[n] \\ & |S|\leq k \end{aligned}$$

How to solve: *Outer Approx. = Dual App + CPM*

3) *Median Regr.:* $\min_{\boldsymbol{\beta}} |r_{(q)}|$, *s.t.* $|r_{(1)}| \leq \dots \leq |r_{(n)}|$

$$\begin{aligned} \min_{\gamma,z_i,\mu_i,\bar{\mu}_i} \ & \gamma \\ \text{s.t.}\ & |r_i| + \bar{\mu}_i \geq \gamma \quad \forall i\in[n] \\ & \gamma \geq |r_i| - \bar{\mu}_i, \quad \forall i\in[n] \\ & M_u z_i \geq \bar{\mu}_i, \quad \forall i\in[n] \\ & M_u(1-z_i) \geq \mu_i, \quad \forall i\in[n] \\ & \sum_{i=1}^{n} z_i = q \\ & \mu_i\geq 0,\ \bar{\mu}_i\geq 0,\ z_i\in\{0,1\} \quad \forall i\in[n] \end{aligned}$$

How to solve: *Sequential Linear Optimization, First-order Subgradient Method*

5. Holistic Regression

When we are performing a linear regression, we would want to observe some desirable properties:

- Sparsity
- Group Sparsity
- Limited Pairwise Multicollinearity
- Detection of relevant nonlinear transformation
- Robustness
- Statistical Significance
- Low Global Multicollinearity

Formulation:

$$\min_{\boldsymbol{\beta},\mathbf{z}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \Gamma\|\boldsymbol{\beta}\|_1$$

$$\begin{aligned} \text{s.t.}\ & \sum_{i=1}^{p} z_i \leq k, \\ & -Mz_i \leq \beta_i \leq Mz_i \quad \forall i\in[p], \\ & z_i = z_j \quad \forall i,j \in GS_m, \forall m\in[G], \\ & z_i + z_j \leq 1, \quad \forall i,j \in \mathcal{HC}_p, \\ & \sum_{i\in\mathcal{T}_j} z_i \leq 1, \quad \forall j\in[p], \\ & \frac{\beta_j}{\tilde{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}} + Mb_j \geq t_\alpha z_j \\ & -\frac{\beta_j}{\tilde{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}} + M(1-b_j) \geq t_\alpha z_j \\ & \sum_{i\in\text{Supp}(\mathbf{a})} \leq |\text{Supp}(\mathbf{a})| - 1, \quad \mathbf{a}\ \text{from Alg. 5.1,} \\ & z_i, b_i \in\{0,1\}, \quad \forall i\in[p] \end{aligned}$$

How to solve: *Cutting plane algorithm, Delayed constraint generation*

6. Sparse and Robust Classification

When we are performing a linear regression, we would want to observe some desirable properties:

Formulation:

$$\min_{\boldsymbol{\beta}\in R^p} \sum_{i=1}^{n}\ell(y_i,\mathbf{x}_i^\top\boldsymbol{\beta}), \quad \text{s.t.}\ g(\boldsymbol{\beta})\leq\delta$$

1) Regularized logistic regression:

$$\min_{\boldsymbol{\beta}\in R^p} \sum_{i=1}^{n}\ell(y_i,\mathbf{x}_i^\top\boldsymbol{\beta}) + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2$$

Under the assumption $\ell(y,\cdot)$ convex for $y\in\{-1,1\}$, we have a close form solution:

$$\max_{\boldsymbol{\alpha}\in R^n} -\sum_{i=1}^{n}\hat{\ell}(y_i,\alpha_i) - \frac{\gamma}{2}\boldsymbol{\alpha}^\top\mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha}, \quad \text{s.t.}\ \boldsymbol{\alpha}^\top\mathbf{e}=0$$

with $\hat{\ell}(y,\alpha) = \max_{u\in R}(u\alpha - \ell(y,u))$ (Fenchel conj.)

2) Sparse Regularized classification:

$$\min_{\boldsymbol{\beta}\in R^p} \sum_{i=1}^{n}\ell(y_i,\mathbf{x}_i^\top\boldsymbol{\beta}) + \frac{1}{2\gamma}\|\boldsymbol{\beta}\|_2^2, \quad \|\boldsymbol{\beta}\|_0\leq k$$

has a dual equivalent problem :
$\min_{\mathbf{s}\in S_k^p} c(\mathbf{s})$ with $S_k^p = \{\mathbf{s}\in\{0,1\}^p, \mathbf{e}^\top\mathbf{s}\leq k\}$

$$c(\mathbf{s}) = max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\sum_{i=1}^{n}\hat{\ell}(y_i,\alpha_i) - \frac{\gamma}{2}\sum_{i=1}^{n} s_j\boldsymbol{\alpha}^\top\mathbf{X}_j\mathbf{X}_j^\top\boldsymbol{\alpha}$$

$$\text{s.t.}\ \mathbf{e}^\top\boldsymbol{\alpha}=0$$

How to solve: *Outer-approximation algorithm*

3) Robust classification:

Include the uncertainty w.r.t. labels:

$$\max_{\boldsymbol{\beta}} \min_{\substack{\Delta\mathbf{y}\in\mathcal{U}_y \\ \Delta\mathbf{X}\in\mathcal{U}_x}} -\sum_{i=1}^{n}\log\left(1+e^{-y_i(1-2\Delta y_i)\boldsymbol{\beta}^\top(\mathbf{x}_i+\Delta\mathbf{x}_i)}\right)$$

has a closed form robust formulation.

13. Optimal Prescription Trees

From Predictions : $\sum_{i=1}^{N} w_{N,i}(\mathbf{x})\mathbf{y}^i$

k-NN : $w_{N,i}(\mathbf{x}) = \begin{cases} \frac{1}{k}\mathbb{1}_{\{\mathbf{x}\in\mathcal{V}_k(\mathbf{x})\}} \\ 0 \quad \text{otherwise} \end{cases}$

CART : $w_{N,i}(\mathbf{x}) = \begin{cases} \frac{1}{|R(\mathbf{x})|}\mathbb{1}_{\{\mathbf{x}\in R(\mathbf{x})\}} \\ 0 \quad \text{otherwise} \end{cases}$

To Prescriptions : $\min_{z\in\mathcal{Z}} \mathbb{E}[c(\mathbf{z},\mathbf{y})|\mathbf{X}=\mathbf{x}]$
$\approx \min_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(\mathbf{x})c(\mathbf{y}^i, z)$

Prescriptions (objectives) :
$\min_{\tau(\mathbf{x})} \mu \left[\sum_{i=1}^{n} \left(y_i \mathbb{1}_{[\tau(\mathbf{x}_i)=z_i]} + \sum_{t\neq z_i} \hat{y}_i(t)\mathbb{1}_{[\tau(\mathbf{x}_i)=t]}\right)\right]$
$+ (1-\mu)\left[\sum_{i=1}^{n}(y_i - \hat{y}_i(z_i))^2\right]$

Strong performances and there are interpretable (Prescriptive analytics).

12. Prescriptive analytics

Thm : if $c(z;y)$ is convex, Z convex, then we can solve the prescription problem in polynomial time
$P = \frac{\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} c(z;y^i) - \sum_{i=1}^{N} c(\hat{z}_N(x^i);y^i)}{\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} c(z;y^i) - \sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c(z;y^i)}$
Measures the prescriptive value of X and of the prescription trained Contrast with R^2

11. Deep Learning and Optimal Trees

An NN is defined by : L hidden layers, 1 output layer
Hidden layer ℓ consisting of N_ℓ nodes;
Some non-linear functions $\phi(x), \phi_0(x)$;
$n_{\ell,i}$ node : $\mathbf{W}_{\ell,i}$, $\mathbf{b}_{\ell,i} \rightarrow y_{\ell,i} = \phi(\mathbf{W}_{\ell,i}^\top \mathbf{y}_{\ell-1} + b_{\ell,i})$ $(\mathbf{y}_0 = \mathbf{x})$

Thm 1 An OCT-H with maximum depth N_1 can classify the data in a training set at least as well as a given classification FNN with the perceptron activation function and N_1 nodes in the first hidden layer. (Conversely)

Thm 2 An OCT-H with maximum depth $q - 1 + \sum_{\ell=1}^{L} N_\ell$ can classify the data in a training set at least as well as a given classification FNN with the ReLU activation function, L hidden layers, N_ℓ nodes. (Conversely)

17. Interpretable Clustering

Objectives : Globally solve *(S)PCA* + OCT to learn clusters.
ICOT algtorithm : Highly interpretable, based on OCT
How to Solve: *Local Search Algorithm*

14. Optimal Design Experiments

Objectives : Randomized experiments are *ineffective* on **small groups** with **high variance** covariate (high prob. to have big differences between groups)
$\mu_p(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n} w_i' x_{ip} \qquad \sigma_p^2(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n}(w_i')^2 x_{ip}$

Formulation :

$$\begin{aligned}
\min_{x,d}\ & d \\
\text{s.t. } & d \geq \mu_p(\mathbf{x}) - \mu_q(\mathbf{x}) + \rho\sigma_p^2(\mathbf{x}) - \rho\sigma_q^2(\mathbf{x}), \ \forall p < q \in [m] \\
& d \geq \mu_p(\mathbf{x}) - \mu_q(\mathbf{x}) + \rho\sigma_q^2(\mathbf{x}) - \rho\sigma_p^2(\mathbf{x}), \ \forall p < q \in [m] \\
& \sum_{i=1}^{n} x_{ip} = k \quad \forall p \in [m], \qquad \sum_{p=1}^{m} x_{ip} = 1 \quad \forall i \in [n] \\
& x_{ip} = 0 \quad \forall i < p \\
& x_{ip} \in \{0,1\}
\end{aligned}$$

Theoretical Results:

$$\mathbb{E}[Z_{ran}] \lesssim \frac{2}{k}\sigma, \quad \mathbb{E}[Z_{opt}(\rho=0)] \lesssim \frac{2\pi}{2^k}\sigma$$

15. Identifying Exceptional Responders

Objectives: Identify a *subpopulation* that would benefits from a treatment, even if the treatment was globally ineffective
Form: $\{\gamma_{s1}, \dots, \gamma_{sK_s}\}$ defines a box-partition of a sub-feature space (*reg.*, *comp.*, *interp.*)

20. Sparse Inverse Covariance Estimation

Objectives : Sparsity is needed.
High Dimensional setting $n \gg p$: sample covariance matrix is **singular**
Parsimonious Model: desirable to have simple model with strong predictive power (sparsity (ℓ_0) > robustness (ℓ_1))
Form. Robustness:
$\min_{\Theta\succ 0}\langle\hat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle - \log\det\mathbf{\Theta} + \|\mathbf{\Theta}\|_1$ **how to solve:** *GLasso Form. Robustness*:
$\min_{\Theta\succ 0}\langle\hat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle - \log\det\mathbf{\Theta}$
$\|\mathbf{\Theta}\|_0 \leq k$ **how to solve:** *Mixed Integer Formulation for sparsity, Big-M Method*

21. Matrix Completion

Objectives: From incomplete matrix $A \in \mathbb{R}^{n\times m}$, n users and m products with existing $Supp(A) = \Sigma$, complete it assuming **low rank hypothesis**.
Form.: $\min_{\|\mathbf{V}\|_2=1}\min_U \frac{1}{n}\left(\sum_{(i,j)\in\Sigma}(X_{ij} - A_{ij})\right)^2 = \frac{1}{\gamma}\|\mathbf{U}\|_2^2$ s.t. $\mathbf{X} = \mathbf{UV}$
Re-Form.: $\min_{\|\mathbf{V}\|_2=1} c(\mathbf{V}) = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{a}}_i^\top\left((\mathbf{I} - \mathbf{V}\left(\frac{\mathbf{I}_k}{\gamma} + \mathbf{V}^\top\mathbf{W}_i\mathbf{V}\right)^{-1}\mathbf{V}^\top\right)\bar{\mathbf{a}}_i$
with
$\mathbf{W}_1, \dots, \mathbf{W}_n \in \mathbb{R}^{m\times m}$ are indicator diagonal matrices : $(\mathbf{W}_i)_{jj} = 1 \equiv \bar{\mathbf{a}}_i = \mathbf{a}_i\mathbf{W}_i$
(allows to keep track of existing values)

Interpretable Matrix Completion:
Objectives: Adding features through $\mathbf{B}$ and insuring sparsity through $\mathbf{S}$

$$\min_{\mathbf{U},\mathbf{S}} \frac{1}{n}\left(\sum_{(i,j)\in\Sigma}(X_{ij} - A_{ij})\right)^2 + R(\mathbf{U},\mathbf{S}) \text{ s.t. } \mathbf{X} = \mathbf{USB}$$

Re-Form.(Int.): $\min_{\mathbf{s}\in S_k^p} c(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{a}}_i^\top\left((\mathbf{I}_m + \gamma\mathbf{W}_i\left(\sum_{j=1}^{p} s_j\mathbf{K}_j\right)\mathbf{W}_i\right)^{-1}\bar{\mathbf{a}}_i$
with
$\mathbf{W}_1, \dots, \mathbf{W}_n \in \mathbb{R}^{m\times m}$ are indicator diagonal matrices : $(\mathbf{W}_i)_{jj} = 1 \equiv \bar{\mathbf{a}}_i = \mathbf{a}_i\mathbf{W}_i$
$\mathbf{K}_j = \mathbf{b}_j\mathbf{b}_j^\top \in \mathbb{R}^{m\times m}$
How to solve: *Stochastic Cutting Planes Methods* (probability of failure decreases exponentially with the number of columns in A)

16. Missing Data Imputation

Type of missing-ness : MCAR (Missing Completely at Random), MAR (Missing at Random), NMAR (Not Missing at Random)
Objectives : Globally impute the missing data (not sequentially) to a set $\mathcal{I}$
`opt.knn`, **Form.** :

$$\begin{aligned}
\min \quad & c(Z,W,X) = \sum_{i\in\mathcal{I}}\sum_{j\neq i} z_{ij}\|w_i - w_j\|_2^2 \\
\text{s.t.} \quad & w_{id} = x_{id}, \quad (i,d) \notin \mathcal{M}, \\
& \textstyle\sum_{j\neq i} z_{ij} = K, \quad i \in \mathcal{I}, \\
& Z \in \{0,1\}^{|\mathcal{I}|\times(n-1)},
\end{aligned}$$

`opt.knn`, **Scal.** : $n \sim 100{,}000's$, $p \sim 1{,}000's$ (Good)
How to Solve: *Block Coordinate Descent, Coordinate Descent*
Further methods: `opt.tree`, `opt.svm`, `opt.cv`

18. Sparse Principal Component Analysis

Objectives : Find a PCA methods that reduces the **noise**, achieved a desired **sparsity**.

Form. : $\max_{\mathbf{x}} \mathbf{x}^{'} \mathbf{\Sigma} \mathbf{x}$ s.t. $\|\mathbf{x}\|_2 = 1,\ \|\mathbf{x}\|_0 \leq k$

How to solve:

First-order methods

$$\min_{\mathbf{x}} \left\| \mathbf{x} - \left(\bar{x} + \frac{1}{L} \nabla f(\bar{\mathbf{x}}) \right) \right\|_2^2 = \|\mathbf{x} - \mathbf{c}\|_2^2$$

$$\|\mathbf{x}\|_2 = 1,\ \|\mathbf{x}\|_0 \leq k$$

$$\mathbf{H}_k(\mathbf{c}) = \begin{cases} \frac{c_i}{\sqrt{\sum_{i \in I} c_i^2}} & i \in \mathbf{I}(\mathbf{c}) \\ 0 & \text{otherwise} \end{cases}$$

with $\mathbf{I}(\mathbf{c}) = \{i_1, \ldots, i_k\}, \quad |c_{i_1}| \geq |c_{i_1}| \geq \ldots \geq |c_{i_n}|$

Binary MIO formulation

Theoretical Bounds

$\mathbf{x}^{'} \mathbf{\Sigma} \mathbf{x} \leq \lambda(\hat{\mathbf{\Sigma}}(X_0))$, $\mathbf{x}^{'} \mathbf{\Sigma} \mathbf{x} \leq M(\hat{\mathbf{\Sigma}}(X_0))$ (Gershgorin)

19. Factor Analysis

Objectives: Obtaining a parsimonious representation of the correlatin structure among a set of variables in terms of a smaller number of common factors.

Model assumptions: $\mathbf{x} = \mathbf{L}\mathbf{f} + \boldsymbol{\epsilon}$

$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{f}] = \mathbb{E}[\boldsymbol{\epsilon}] = 0$

$\mathrm{Cov}[\boldsymbol{\epsilon}] = \mathbf{\Psi} = \mathrm{diag}(\Psi_1, \ldots, \Psi_p)$, $\mathrm{Cov}[\mathbf{f}, \boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Cov}[\mathbf{f}] = \mathbf{I}$

$$\mathbf{\Sigma} = \mathbf{\Sigma}_c + \mathbf{\Psi} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}$$

Low rank assumption (Parsimonious Model):

$$\mathbf{\Sigma} = \mathbf{\Theta} + \boldsymbol{\mathcal{N}} + \mathbf{\Psi} = \mathbf{L}_1\mathbf{L}_1^\top + (\mathbf{\Sigma}_c - \mathbf{L}_1\mathbf{L}_1^\top) + \mathbf{\Psi}$$

Formulation:

$$\begin{aligned} \min \quad & \eta_{\mathbf{\Sigma}}(\mathbf{\Theta}, \mathbf{\Psi}) := \|\mathbf{\Sigma} - (\mathbf{\Theta} + \mathbf{\Psi})\|_q^q \\ & \text{s.t.}\ \mathrm{rank}(\mathbf{\Theta}) \leq r, \quad \mathbf{\Theta} \succcurlyeq 0 \\ & \mathbf{\Psi} = \mathrm{diag}(\Psi_1, \ldots, \Psi_p) \succcurlyeq 0 \\ & \mathbf{\Sigma} - \mathbf{\Psi} \succcurlyeq 0 \end{aligned}$$

How to solve: *Conditional Gradient Algorithm for Smooth Problems, digression: Conditional Gradient Algorithm, Concave Gradient Descent*

Remarks: Roughly scalable, we can prove optimality using branch and bound

10(a). Optimal Classification Trees

Decision Tree: $\min error(\mathbb{T}, X, y) + \alpha \times complexity(\mathbb{T})$

OCT with Parallel Splits

Terminology: $\mathcal{T}_B = \{1, \ldots, \lfloor \frac{T}{2} \rfloor\}$, $\mathcal{T}_B = \{\lfloor \frac{T}{2} \rfloor + 1, \ldots, T\}$ branch nodes and lead nodes

Input Conditions: $\mathbf{x}_i \in [0,1]^p$, $y_i \in [K], \forall t \in \mathcal{T}_B$

Constraints:

$$\begin{aligned} \sum_{j=1}^{p} a_{jt} = d_t, &\quad \forall t \in \mathcal{T}_B \\ 0 \leq b_t \leq d_t, &\quad \forall t \in \mathcal{T}_B \\ d_t \leq d_{p(t)} &\quad \forall t \in \mathcal{T}_B \setminus \{1\} \\ z_{it} \leq l_t &\quad \forall t \in \mathcal{T}_L \\ \sum_{t=1}^{n} z_{it} \geq N_{min} l_t &\quad \forall t \in \mathcal{T}_L \\ \sum_{t \in \mathcal{T}_L} z_{it} = 1 &\quad i \in [n] \end{aligned}$$

$$\mathbf{a}_m^\top (\boldsymbol{\epsilon} - \epsilon_{min} + \mathbf{x}_i) + \epsilon_{min} \leq b_m + (1 + \epsilon_{max})(1 - z_{it}),$$
$$\forall i \in [n], t \in \mathcal{T}_L, m \in \mathcal{L}(t),$$
$$\mathbf{a}_m^\top \mathbf{x}_i \geq b_m - 1(1 - z_{it}), \forall i \in [n], t \in \mathcal{T}_L, m \in \mathcal{R}(t),$$
$$N_{kt} = \sum_{i: y_i = k} z_{it}, \quad \forall t \in \mathcal{T}_L, k \in [K]$$
$$N_t = \sum_i z_{it}$$
$$L_t \geq N_t - N_{kt} - n(1 - c_{kt})$$
$$L_t \leq N_t - N_{kt} + n c_{kt}$$
$$L_t \geq 0$$
$$\sum_{k=1}^{K} c_{kt} = l_t,$$
$$c_{kt} \in \{0,1\}, \mathbf{a} \in \{0,1\}^p, \forall t \in \mathcal{T}_B$$

with :

$$\epsilon_{min} = \min_j \left\{ \min \left\{ x_j^{(i+1)} - x_j^{(i)} \,\middle|\, x_j^{(i+1)} \neq x_j^{(i)},\ i < n \right\} \right\}$$

$$\epsilon_{max} = \max_j \left\{ \min \left\{ x_j^{(i+1)} - x_j^{(i)} \,\middle|\, x_j^{(i+1)} \neq x_j^{(i)},\ i < n \right\} \right\}$$

Objective: $\boxed{\min \sum_{t \in \mathcal{L}} L_t + \alpha \sum_{t \in \mathcal{B}} d_t}$

How to solve: *MIO w. Warm starts, Local search*

OCT with Hyperplane Splits works almost the same (Complexity-wise)

10(b). Optimal Regression Trees

ORT with Constant Predictions

Conditions: $\mathbf{x}_i \in [0,1]^p$, $y_i \in [K], \mathbf{a} \in \{0,1\}^p, \forall t \in \mathcal{T}_B$

Constraints:

$$\begin{aligned} \sum_{j=1}^{p} a_{jt} = d_t, &\quad \forall t \in \mathcal{T}_B \\ 0 \leq b_t \leq d_t, &\quad \forall t \in \mathcal{T}_B \\ d_t \leq d_{p(t)} &\quad \forall t \in \mathcal{T}_B \setminus \{1\} \\ z_{it} \leq l_t &\quad \forall t \in \mathcal{T}_L \\ \sum_{t=1}^{n} z_{it} \geq N_{min} l_t &\quad \forall t \in \mathcal{T}_L \\ \sum_{t \in \mathcal{T}_L} z_{it} = 1 &\quad i \in [n] \end{aligned}$$

$$\mathbf{a}_m^\top (\boldsymbol{\epsilon} - \epsilon_{min} + \mathbf{x}_i) + \epsilon_{min} \leq b_m + (1 + \epsilon_{max})(1 - z_{it}),$$
$$\forall i \in [n], t \in \mathcal{T}_L, m \in \mathcal{L}(t),$$
$$\mathbf{a}_m^\top \mathbf{x}_i \geq b_m - 1(1 - z_{it}), \forall i \in [n], t \in \mathcal{T}_L, m \in \mathcal{R}(t),$$
$$L_i \geq (f_i - y_i)^2$$
$$f_i - \beta_{0t} \geq -M_f(1 - z_{ik}) \quad \forall i \in [n]$$
$$f_i - \beta_{0t} \geq +M_f(1 - z_{ik}) \quad \forall i \in [n]$$
$$L_t \geq 0$$
$$\sum_{k=1}^{K} c_{kt} = l_t$$
$$c_{kt} \in \{0,1\}$$

with :

$$\epsilon_{min} = \min_j \left\{ \min \left\{ x_j^{(i+1)} - x_j^{(i)} \,\middle|\, x_j^{(i+1)} \neq x_j^{(i)},\ i < n \right\} \right\}$$

$$\epsilon_{max} = \max_j \left\{ \min \left\{ x_j^{(i+1)} - x_j^{(i)} \,\middle|\, x_j^{(i+1)} \neq x_j^{(i)},\ i < n \right\} \right\}$$

Objective: $\boxed{\min \sum_{t \in \mathcal{L}} L_t + \alpha \sum_{t \in \mathcal{B}} d_t}$

How to solve: *MIO w. Warm starts, Local search*

ORT with Linear Predictions works similarity

Obj.: $\boxed{\min \sum_{t \in \mathcal{L}} L_t + \alpha \sum_{t \in \mathcal{B}} d_t + \lambda \sum_{t \in \mathcal{T}_L} \sum_{j=1}^{p} r_{jt}}$

Updated Constraints:

$$f_i - (\boldsymbol{\beta}_t^\top \mathbf{x}_i + \beta_{0t}) \geq -M_f(1 - z_{ik}) \quad \forall i \in [n], t \in \mathcal{T}_L$$
$$f_i - (\boldsymbol{\beta}_t^\top \mathbf{x}_i + \beta_{0t}) \leq +M_f(1 - z_{ik}) \quad \forall i \in [n], t \in \mathcal{T}_L$$
$$-M_r r_{jt} \leq \beta_{jt} \leq M_r r_{jt}, \forall j \in [p], \forall t \in \mathcal{T}_L$$