A Block Coordinate Gradient Descent Method for Regularized Convex Separable Optimization and Covariance Selection

Sangwoon Yun

Computational Sciences
Korea Institute for Advanced Study

April 30, 2010
Ewha Womans University
(Joint work with Paul Tseng (UW) and Kim-Chuan Toh (NUS))

Outline

- Motivation: Covariance Selection
- General Problem Model:
  Regularized Convex Separable Optimization
- Block Coordinate Gradient Descent Method
- Convergence Results
- Numerical Experience
- Conclusions & Future Work

Motivation: Covariance Selection

Given m i.i.d. observations $x^{(1)}, ..., x^{(m)}$ drawn from a n-dimensional Gaussian distribution $N(x; \mu; \Sigma)$, sample covariance matrix $\hat{\Sigma}$ is defined as

$$\hat{\Sigma} := \frac{1}{m} \sum_{k=1}^{m} (x^{(k)} - \hat{\mu})(x^{(k)} - \hat{\mu})^T,$$

where $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$ is the sample mean.

If Σ is non-singular, then the probability density function of x is given by

$$P(x; \mu; \Sigma) = \frac{1}{(2\pi)^{n/2} (\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (x - \hat{\mu})^T \Sigma^{-1} (x - \hat{\mu}) \right).$$

To estimate Σ from the sample $\mathcal{X} := \{x^{(1)}, ..., x^{(m)}\}$, consider the log-likelihood function

$$\log P(\mathcal{X}; \mu; \Sigma) = -\frac{m}{2} \log(\det\Sigma) - \frac{1}{2} \sum_{k=1}^{m} (x^{(k)} - \hat{\mu})^T \Sigma^{-1} (x^{(k)} - \hat{\mu}) + c,$$

where c is a constant.

By using the inner product $\langle X,\ Y\rangle = \mathrm{Trace}(XY)$ for $X, Y \in \mathcal{S}^n$, the previous expression can be rewritten as

$$\log P(\mathcal{X}; \mu; \Sigma) = \frac{m}{2} \log(\det \Sigma^{-1}) - \frac{m}{2} \left\langle \Sigma^{-1},\ \hat{\Sigma} \right\rangle + c$$

If $\hat{\Sigma}$ is nonsingular (hence $m \geq n$), then
$\hat{\Sigma}^{-1} = \arg\max\{\log P(\mathcal{X}; \mu; \Sigma) | \Sigma \in \mathcal{S}^n_{++}\}$ is the maximum likelihood estimator of Σ^{-1}.

In practice, one may not want to use $\hat{\Sigma}^{-1}$ as the estimator of Σ^{-1}.

- $\hat{\Sigma}$ is singular or nearly so, it is not a robust estimator of Σ^{-1} for many statistical purposes.
- May want to impose structural conditions on Σ^{-1}, such as conditional independence between different components of x, which is reflected as zero entries in Σ^{-1}.

- Covariance selection problem was first introduced by Dempster (1972), who suggested that the covariance structure of a multivariate normal population can be simplified by setting elements of the inverse covariance matrix to zero.
- Since then, covariance selection model has become a common statistical tool to distinguish direct from indirect interactions among a set of variables.
- Gaussian Graphical Model (GGM) (96' Laurizen, 00' Edwards) is the graphical interpretation of covariance selection model and is another popular tool .

Applications of covariance selection model can be found in various areas.

- In financial portfolio management (a sparse portfolio can be generated by the covariance selection model)
- In the research of dependency networks of genome data (sparse gene association network exhibited in GGM can help to explain the known biological pathways and to provide insights on the unknowns)

  Recent advances in DNA microarray technology have led to the challenge problem of modeling associations for a large number of genes (say, $10^3 - 10^4$) from a small number of available samples (say, 10^2).

In such an application, the sample covariance matrix $\hat{\Sigma}$ is singular.

Covariance selection problems can be modeled as log-det semidefinite programming (SDP) problems

- If no sparsity pattern is assumed,
  - Estimation of the sparsity pattern can be achieved by ℓ_1-regularized maximum log-likelihood estimation:

$$\max_{X \in \mathcal{S}^n} \quad \log\det X - \left\langle \hat{\Sigma},\, X \right\rangle - \sum_{i,j=1}^{n} \rho_{ij} |X_{ij}|,$$

  - $\rho_{ij} > 0$: parameter controlling the trade-off between the goodness-of-fit and the sparsity of X.
  - $-\log\det X + \left\langle \hat{\Sigma},\, X \right\rangle$ is strictly convex, cont. diff. on its domain $\mathcal{S}^n_{++}$, $O(n^3)$ opers. to evaluate. $\sum_{i,j=1}^{n} \rho_{ij}|X_{ij}|$ is convex, nonsmooth.
    In applications, n can exceed 5000.
  - The dual problem can be formulated as follows:

$$\begin{array}{ll} \min\limits_{X \in \mathcal{S}^n} & -\log\det X - n \\ \text{s.t.} & |(X - \hat{\Sigma})_{ij}| \le \rho_{ij},\ i,j = 1, \ldots, n. \end{array}$$

- If conditional independence structure between all the variables are given,
  - Covariance selection problem can be formulated as a log-det maximization problem with linear constraints, that is, finding the maximum log-likelihood value subject to given entry-wise constraints:

$$\begin{aligned} \max_{X \in \mathcal{S}^n} \quad & \log \det X - \left\langle \hat{\Sigma},\, X \right\rangle \\ \text{s.t.} \quad & X_{ij} = 0,\ \forall (i,j) \in V, \end{aligned}$$

    where V is a collection of all pairs of conditional independent nodes. We note that $(i,i) \notin V$ for $1 \leq i \leq n$ and $(i,j) \in V$ if and only if $(j,i) \in V$.

- Previous primal problems can be considered as special cases of the following more general log-det semidefinite programming problem:

$$\begin{array}{ll} \max\limits_{X \in \mathcal{S}^n} & \log\det X - \left\langle \hat{\Sigma},\, X \right\rangle - \sum_{(i,j)\notin V} \rho_{ij}|X_{ij}| \\ \text{s.t.} & X_{ij} = 0,\ \forall (i,j) \in V, \end{array}$$

- The dual problem can be expressed:

$$\begin{array}{ll} \min\limits_{X \in \mathcal{S}^n} & -\log\det X - n \\ \text{s.t.} & |(X - \hat{\Sigma})_{ij}| \le \upsilon_{ij},\ i,j = 1,...,n, \end{array}$$

  where $\upsilon_{ij} = \rho_{ij}$ for all $(i,j) \notin V$ and $\upsilon_{ij} = \infty$ for all $(i,j) \in V$.

- Can in principle be solved by popular interior-point method based solvers such as SDPT3 or SeDuMi.
- Resulting log-det SDP problems typically have large number of linear constraints p (even for moderate n, say $n \le 100$)
- Solvers (SDPT3 or SeDuMi) cannot handle since the computational cost in each iteration is at least $O(p^3)$ and the memory required is at least $O(p^2)$ bytes.

Recent Algorithms

- Unconstrained Problems (no given sparsity pattern)
  - Nesterov's smooth gradient method (d'Aspremont '08)
  - Block coordinate descent method (d'Aspremont '08, Friedman '08)

    Use coord. des. method to solve dual problem, cycling thru columns (& rows) of X. Each iter. reduces (via determinant property) to

$$\begin{aligned} \min_{x \in \Re^{n-1}} \quad & x^T X_{j^c j^c}^{-1} x \\ \text{s.t.} \quad & |x - \hat{\Sigma}_{j^c j}| \leq \rho_{j^c j}, \end{aligned}$$

    (Solve this using IP method $O(n^3)$ opers. or coordinate descent method) or (via determinant property & duality) to

$$\min_{x \in \Re^{n-1}} \frac{1}{2} x^T X_{j^c j^c} x - \hat{\Sigma}_{j^c j}^T x + \rho_{j^c j}^T |x|.$$

    (Solve this using coordinate descent method).

- Greedy algorithm (based on a coordinate ascent method) (Scheinberg '09)
- Alternating direction method (Yuan '09)

- Constrained Problems:
  - Newton method (PCG) (Dahl '08)
  - (Adaptive) Nesterov's smooth method (Lu '09)
    solving a sequence of penalized problems of the primal form.
  - Semismooth Newton-CG method (PCG) (Wang '09)

General Problem Model: Regularized Convex Separable Optimization

$$\min_{X \in \Re^{m \times n}} \quad F(X) := f(X) + P(X),$$

- f: real-valued, convex smooth on $\mathrm{dom} f$
- P: proper, convex, lsc
- P: separable i.e., $P(X) = \sum_{i,j=1}^{n} P_{ij}(X_{ij})$

- for primal covariance selection problem,

$$\begin{aligned} f(X) &= -\log\det X + \left\langle \hat{\Sigma},\, X \right\rangle \\ P_{ij}(X_{ij}) &= \rho_{ij}|X_{ij}| \;\forall (i,j) \notin V, \quad P_{ij} \equiv \delta_{\{0\}} \;\forall (i,j) \in V, \quad X \in \mathcal{S}^n. \end{aligned}$$

- for dual covariance selection problem,

$$\begin{aligned} f(X) &= -\log\det X - n \\ P_{ij}(X_{ij}) &= \begin{cases} 0 & \text{if } -U_{ij} \le (X-S)_{ij} \le U_{ij}; \\ \infty & \text{else,} \end{cases}, \quad X \in \mathcal{S}^n, \end{aligned}$$

where $U_{ij} = \rho_{ij}$ for all $(i,j) \notin V$, and $U_{ij} = \infty$ for all $(i,j) \in V$.

Block Coordinate Gradient Descent Method

Descent direction

For $X \in \mathrm{dom} F$, choose $\mathcal{J}(\neq \emptyset) \subseteq \mathcal{N} = \{11, 12, ..., nn\}$ and a self-adjoint p.d. linear map. $\mathcal{H}$, Then solve

$$\min_{D \in \Re^{m \times n}, D_{ij}=0, \forall (i,j) \notin \mathcal{J}} \{\langle \nabla f(X),\, D\rangle + \frac{1}{2}\langle D,\, \mathcal{H}(D)\rangle + P(X+D) - P(X)\}$$

direc. subprob

Let $D_{\mathcal{H}}(X; \mathcal{J})$ and $q_{\mathcal{H}}(X; \mathcal{J})$ be the opt. soln & obj. value of the direc. subprob.

Properties:

- $D_{\mathcal{H}}(X;\mathcal{N}) = 0 \;\Leftrightarrow F'(X;D) \geq 0 \;\forall D \in \Re^{m\times n}$. stationarity
- if $X \in \mathcal{S}^n$ and $\mathcal{H}(D) = (H_{ij}D_{ij})_{ij}$ where $H \in \mathcal{S}^n$ with $H_{ij} > 0 \;\Rightarrow$
$D_{\mathcal{H}}(X;\mathcal{J}) = \sum_{(i,j)\in\mathcal{J}} D_{\mathcal{H}}(X;(i,j))$, $q_{\mathcal{H}}(X;\mathcal{J}) = \sum_{(i,j)\in\mathcal{J}} q_{\mathcal{H}}(X;(i,j))$. separab.
  - If $P_{ij} \equiv 0$, then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = -\frac{(\nabla f(X))_{ij}}{H_{ij}}$.
  - If P_{ij} is an indicator function of the bounded constraint (i.e., $-U_{ij} \leq X_{ij} \leq U_{ij}$),
    then $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = \text{median}\left\{-U_{ij} - X_{ij}, -\frac{(\nabla f(X))_{ij}}{H_{ij}}, U_{ij} - X_{ij}\right\}$.
  - If $P_{ij}(X) = \rho_{ij}|X_{ij}|$, then
    $(D_{\mathcal{H}}(X;\mathcal{N}))_{ij} = -\text{median}\left\{\frac{(\nabla f(X))_{ij}-\rho_{ij}}{H_{ij}}, X_{ij}, \frac{(\nabla f(X))_{ij}+\rho_{ij}}{H_{ij}}\right\}$.
- $q_{\mathcal{H}}(X;\mathcal{J}) \leq -\frac{1}{2}\langle D, \mathcal{H}(D)\rangle$ where $D = D_{\mathcal{H}}(X;\mathcal{J})$.

Stepsize: **Armijo rule**

Choose α to be the largest element of $\{\beta^k\}_{k=0,1,\ldots}$ satisfying $F(X + \alpha D) \leq F(X) + \alpha\sigma q_{\mathcal{H}}(X; \mathcal{J})$ $(0 < \beta < 1, 0 < \sigma < 1)$.

For covariance selection problems, the limited minimization rule

$$\alpha \in \arg\min_t \{F(X + tD) \mid 0 \leq t \leq s\},$$

where $0 < s < \infty$, can also be used.

Choose $\mathcal{J}$:

- Gauss-Seidel: $\mathcal{J}^0, \mathcal{J}^1, \ldots$ collectively cover $\mathcal{N}$ for every T consecutive iterations, where $T \geq 1$

$$\mathcal{J}^k \cup \mathcal{J}^{k+1} \cup \cdots \cup \mathcal{J}^{k+T-1} = \mathcal{N}, \quad k = 0, 1, \ldots$$

- Restricted Gauss-Seidel: there exists a subsequence $\mathcal{T} \subseteq \{0, 1, \ldots\}$ such that

$$0 \in \mathcal{T}, \qquad \mathcal{N} = \left(\text{disjoint union of } \mathcal{J}^k, \mathcal{J}^{k+1}, \ldots, \mathcal{J}^{\tau(k)-1}\right) \quad \forall k \in \mathcal{T},$$

where $\tau(k) := \min\{k' \in \mathcal{T} \mid k' > k\}$.

For the covariance selection problems,

- $f(X) = -\log\det X + \left\langle \hat{\Sigma},\, X \right\rangle$ with $X \in \mathcal{S}^n$ and $\hat{\Sigma} \succeq 0_n$.
- want to avoid computing $\det(X + \alpha D)$ and $\nabla f(X) = X^{-1} + \hat{\Sigma}$ from scratch since it would require $O(n^3)$ opers.
- For Gauss-Seidel rule, choose $\mathcal{J}^k = \{(i,j),(j,i) \mid i = 1,...,n\}$ where $j = k + 1 (\mathrm{mod}\ n)$
- For restricted Gauss-Seidel rule, choose $\mathcal{J}^k = \{(i,j),(j,i) \mid i = 1,...,j\}$ where $j = k + 1 (\mathrm{mod}\ n)$
- Update only one column (and corresponding row) of X at each iteration.
- $\det(X + \alpha D)$ can be computed in $O(n^2)$ opers. by using the Schur complement of $(X + \alpha D)_{j^c j^c}$.
- $(X^{\mathrm{new}})^{-1}$ can be updated in $O(n^2)$ operations from X^{-1} using the Sherman-Woodbury-Morrison formula.

Convergence Results

Global convergence If

- $\underline{\lambda} \leq \lambda_{\min}(\mathcal{H})$ and $\lambda_{\max}(\mathcal{H}) \leq \bar{\lambda}$, where $0 < \underline{\lambda} \leq \bar{\lambda}$
- α is chosen by Armijo rule

then every cluster point of the X-sequence generated by BCGD method using Gauss-Seidel rule is a stationary point of F.

Assumption
(a) There exists a positive constant Λ such that

$$\|X\| \leq \Lambda \ \forall X \in \mathcal{X}^0 := \{X \mid F(X) \leq F(X^0)\}.$$

(b) There exist $L \geq 0$, $\varrho > 0$ such that $\mathcal{X}^0_\varrho := \left(\mathcal{X}^0 + \varrho B\right) \cap \mathrm{dom} P \subseteq \mathrm{dom} f$ and

$$\|\nabla f(Y) - \nabla f(Z)\| \leq L\|Y - Z\| \quad \forall Y, Z \in \mathcal{X}^0_\varrho.$$

Local convergence rate If

- $\underline{\lambda} \leq \lambda_{\min}(\mathcal{H})$ and $\lambda_{\max}(\mathcal{H}) \leq \bar{\lambda}$, where $0 < \underline{\lambda} \leq \bar{\lambda}$
- α is chosen by Armijo rule
- Assumption (a) is satisfied
- C1: f is strongly convex, and Assumption (b) is satisfied

  C2: $f(X) = g(\mathcal{E}(X)) + \langle Q, X \rangle$ for all $X \in \Re^{m \times n}$, where $\mathcal{E} : \Re^{m \times n} \to \Re^{m \times n}$ is a linear mapping, $Q \in \Re^{m \times n}$, and g is differentiable on $\mathrm{dom} g$, g is strongly convex and Assumption (b) is satisfied with replacing f by g and $\mathcal{X}^0$ by any level set of g. P is polyhedral.

then the X-sequence generated by BCGD method using restricted Gauss-Seidel rule converges at R-linear rate.

Iteration Complexity If

- $\underline{\lambda} \leq \lambda_{\min}(\mathcal{H}^k)$ and $\lambda_{\max}(\mathcal{H}^k) \leq \bar{\lambda}$ for all k, where $0 < \underline{\lambda} \leq \bar{\lambda}$.
- α^k is chosen by Armijo rule with $\inf_k \alpha^k_{\text{init}} > 0$ and $\sup_k \alpha^k < \infty$
- Assumption (a) and (b) are satisfied, $X^k + \min\{\alpha^k_{\text{init}}, \frac{\alpha^k}{\beta}\} D^k \in \mathcal{X}^0_\varrho \ \forall k$
- $\inf_X F(X) > -\infty$, $\{\mathcal{J}^k\}$ is chosen by Gauss-Seidel rule

If $\sqrt{e^k - e^{\tau(k)}} \geq C_2/C_1, \ \forall\ k > 0$, then $e^k \leq \epsilon$ whenever

$$t_k \ \geq \ \left\lceil \ln(e^0/\epsilon)/\ln(1 - 1/(2C_1))^{-1} \right\rceil;$$

otherwise, $e^k \leq \epsilon$ whenever

$$t_k \geq \left\lceil 4C_2^2/\epsilon \right\rceil.$$

- $e^k := F(X^k) - \inf_X F(X)$, $r^0 := \max_X \left\{ \text{dist}(X, \bar{\mathcal{X}}) \mid X \in \mathcal{X}^0 \right\}$
- $t_k = \max\{i \mid \tau^i(0) \leq k\}$, $C_1 = \frac{2C_0}{\sigma}$, $C_2 = \sqrt{2\bar{\lambda}C_1}\, r^0$
- $C_0 = \frac{1}{\underline{\alpha}} + \frac{\sup_\ell \alpha^\ell \max_\ell \{\sum_{i=\ell}^{\tau(\ell)-1} L_i\}}{\underline{\lambda}^2}$ (L_i: Lipschitz constant of $\nabla f(\cdot)_{\mathcal{J}^i}$ over $\mathcal{X}^0$)
- $\underline{\alpha} := \min\{\inf_k \alpha^k_{\text{init}}, \beta \min\{1, 2\underline{\lambda}(1-\sigma)/L, \varrho/\sup_\ell \|D^\ell\|\}\}$.

Iteration complexity for covariance selection

- Assumption (a) (Boundedness):

  Theorem: Suppose $P(X) \geq \varpi \mathrm{tr}(X)$ whenever $X \succ 0$ and $\mathrm{tr}(X) \geq \zeta$, for some scalars $\varpi > 0$, $\zeta > 0$. Then there exist positive constants $\underline{\zeta}$ and $\bar{\zeta}$ such that $\underline{\zeta} I \preceq X \preceq \bar{\zeta} I$ for all $X \in \mathcal{X}^0$.

  - if $P_{ij}(X_{ij}) = \begin{cases} \rho_{ij}|X_{ij}| & \text{if } (i,j) \notin V; \\ \delta_{\{0\}} & \text{else,} \end{cases}$, $\zeta > 0$, and $\varpi = \min_i\{\rho_{ii}\}$ with $\rho_{ii} > 0$ for $i = 1, ..., n$.
  - if $P_{ij}(X_{ij}) = \begin{cases} 0 & \text{if } -U_{ij} \leq (X - S)_{ij} \leq U_{ij}; \\ \infty & \text{else,} \end{cases}$, where $U_{ij} = \rho_{ij}\ \forall (i,j) \notin V$ and $U_{ij} = \infty\ \forall (i,j) \in V$, $\varpi > 0$, and $\zeta > \mathrm{tr}(\hat{\Sigma} + U)$ with $U_{ii} < \infty$ for $i = 1, ..., n$.

- Assumption (b) (Lipschitz continuity) :

  $L = 1/((1-\omega)^2 \underline{\zeta}^2)$, $\varrho = \omega \underline{\zeta}$ for $0 < \omega < 1$

For the dual problems, if

- $\mathcal{J}^k = \{(i,j),(j,i) \mid i = 1,...,n\}$, where $j = k+1 (\mathrm{mod}\ n)$
- $\mathcal{H}^k = I$, $\omega = \frac{1}{2}$

Then iteration bounds for achieving ϵ-optimality:

$$O\left(\frac{n}{\underline{\zeta}^2} \ln\left(\frac{e^0}{\epsilon}\right)\right) \quad \text{or} \quad O\left(\frac{n^2\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right).$$

Since $\mathcal{N}$ is the union of $\mathcal{J}^k, \mathcal{J}^{k+1}, \ldots, \mathcal{J}^{k+n-1}$, the resulting complexity bounds on the number of iterations for achieving ϵ-optimality can be

$$O\left(\frac{n^2}{\underline{\zeta}^2} \ln\left(\frac{e^0}{\epsilon}\right)\right) \quad \text{or} \quad O\left(\frac{n^3\bar{\zeta}^2}{\epsilon\underline{\zeta}^2}\right).$$

- Computational cost per each iteration of BCGD method is $O(n^2)$ operations except the first iteration
- BCGD method can be implemented to achieve ϵ-optimality in

$$O\left(\frac{n^5}{\epsilon}\right)$$

operations.
- Worst-case arithmetic cost of the first-order method proposed by Lu to achieve ϵ-optimality for unconstrained problem is $O(n^4/\sqrt{\epsilon})$ operations.

Numerical Experience

The dual form of the covariance selection:

$$\begin{aligned} \min_{X \in \mathcal{S}^n} \quad & -\log\det X - n \\ \text{s.t.} \quad & |(X - \hat{\Sigma})_{ij}| \leq \upsilon_{ij}, \ i,j = 1,...,n, \end{aligned}$$

where $\upsilon_{ij} = \rho_{ij}$ for all $(i,j) \notin V$ and $\upsilon_{ij} = \infty$ for all $(i,j) \in V$.

- Implement BCGD method in Matlab.
- Choose $\mathcal{H}^k$ to satisfy $\mathcal{H}^k(D) = (H^k_{ij} D_{ij})_{ij}$, where $H^k = h^k (h^k)^T$ with $h^k_j = \min\{\max\{((X^k)^{-1})_{jj}, 10^{-10}\}, 10^{10}\}\ \forall j = 1,...,n$.

  If $10^{-10} \leq ((X^k)^{-1})_{jj} \leq 10^{10}$ for all $j = 1,...,n$, then this choice can be viewed as a diagonal approximation to the Hessian.
- Choose $\mathcal{J}$ by Gauss-Seidel rule, $\mathcal{J}^k = \{(i,j),(j,i) \mid i = 1,...,n\}$ where $j = k + 1 (\mathrm{mod}\ n)$.

  Update only one column (and corresponding row) at each iteration.

- Choose α^k by the limited minimization rule.

$$\alpha^k \in \underset{0 \le \alpha \le s}{\arg\min}\{-\log\det(X^k + \alpha D^k) - n \mid |(X^k + \alpha D^k - \hat{\Sigma})_{ij}| \le \upsilon_{ij}\},$$

  - By permutation

$$X^k = \begin{pmatrix} V^k & u^k \\ (u^k)^T & w^k \end{pmatrix} \text{ and } D^k = \begin{pmatrix} 0_{n-1} & d^k \\ (d^k)^T & r^k \end{pmatrix},$$

    where $V^k \in \mathcal{S}^{n-1}$, $u^k,\ d^k \in \Re^{n-1}$, and $w^k,\ r^k \in \Re$.

  - $X^k + \alpha D^k \succ 0$ iff $V^k \succ 0$ and $w^k + \alpha r^k - (u^k + \alpha d^k)^T (V^k)^{-1}(u^k + \alpha d^k) > 0$.

  - Since $X^k \succ 0$, $V^k \succ 0$, $X^k + \alpha D^k \succ 0$ iff

$$a_1^k \alpha^2 + 2a_2^k \alpha - a_3^k < 0,$$

    where $a_1^k = (d^k)^T (V^k)^{-1}(d^k)$, $a_2^k = (u^k)^T (V^k)^{-1}(d^k) - 0.5 r^k$ and $a_3^k = w^k - (u^k)^T (V^k)^{-1}(u^k) > 0$.

- The difference of objective values

$$\begin{aligned} & -\log\det(X^k+\alpha D^k)+\log\det(X^k) \\ = \quad & -\log\det V^k-\log\left(w^k+\alpha r^k-(u^k+\alpha d^k)^T(V^k)^{-1}(u^k+\alpha d^k)\right) \\ & +\log\det V^k+\log\left(w^k-(u^k)^T(V^k)^{-1}(u^k)\right) \\ = \quad & -\log\left(a_3^k-a_1^k\alpha^2-2a_2^k\alpha\right)+\log a_3^k. \end{aligned}$$

- If $-\log\left(a_3^k-a_1^k\alpha^2-2a_2^k\alpha\right)+\log a_3^k<0$, then $a_1^k\alpha^2+2a_2^k\alpha<0$.

- Quantity $-\log\det(X^k+\alpha D^k)+\log\det(X^k)$ is minimized when

$$\alpha=\begin{cases}\min\{1,-a_2^k/a_1^k\} & \text{if } d^k\neq 0;\\ 1 & \text{else.}\end{cases}$$

- Termination Criterion:

$$\sqrt{\langle D_{H^k}(X^k;\mathcal{N}),\,\mathcal{H}^k(D_{H^k}(X^k;\mathcal{N}))\rangle}\le 5\times 10^{-3},$$

$$\frac{\left|\langle S,(X^k)^{-1}\rangle+\sum_{(i,j)\notin V}\rho_{ij}|((X^k)^{-1})_{ij}|-n\right|}{1+\left|\log\det(X^k)+\langle S,(X^k)^{-1}\rangle+\sum_{(i,j)\notin V}\rho_{ij}|((X^k)^{-1})_{ij}|\right|}\le 10^{-4}.$$

Generating test problems:

- Generate a random sparse matrix $\mathtt{A} \in \mathcal{S}^n$ whose nonzero elements are set randomly to be ± 1.
- Generate a sparse inverse covariance matrix Σ^{-1} from $\mathtt{A}$ as follows:

$$\begin{aligned}
&\mathtt{A} = \mathtt{A} * \mathtt{A}'; \ \mathtt{d} = \mathtt{diag}(\mathtt{A});\\
&T = \mathtt{diag}(\mathtt{d}) + \mathtt{max}(\mathtt{min}(\mathtt{A} - \mathtt{diag}(\mathtt{d}), 1), -1);\\
&\Sigma^{-1} = T - \min\{1.2\lambda_{\min}(T) - 10^{-4}, 0\} I.
\end{aligned}$$

- Generate a matrix $B \in \mathcal{S}^n$ by

$$B = \Sigma + 0.15 \frac{\|\Sigma\|_F}{\|\Xi\|_F} \Xi,$$

  where $\Xi \in \mathcal{S}^n$ is a random matrix whose elements are drawn from the uniform distribution on the interval $[-1, 1]$.
- Finally, obtain randomly generated sample covariance matrix:

$$\hat{\Sigma} = B - \min\{\lambda_{\min}(B) - 10^{-4}, 0\} I.$$

- $\Omega = \{(i,j) \mid (\Sigma^{-1})_{ij} = 0, |i-j| \geq 2\}$.
- For the constraint problem, set V to be a random subset of Ω such that **card**(V) is about 50% of **card**(Ω).
- $\rho_{ij} = 5/n$ for all $(i,j) \notin V$.

The purpose of solving the covariance selection problems

- not to recover the true matrix Σ^{-1} accurately
- but to detect the sparsity pattern of Σ^{-1} while maintaining a reasonable approximation to the true matrix.

The quality of the approximation of Σ^{-1} by X is measured by

$$\mathrm{L}_Q := \frac{1}{n}\|\Sigma X - I\|_F$$

The quality of sparsity pattern is measured by $\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$ and $\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$

TP, TN, FP, and FN denotes the number of true positives, true negatives, false positives, and false negatives, respectively, with respect to the sparsity pattern of Σ^{-1}.

Test Results

Estimated matrix X would not be sparse in general but have many small entries.
Postprocess the matrix X by setting all entries which are smaller than 5×10^{-2} in absolute value to 0.

n | density(%) | **card**(V)	iteration count		primal objective value		time (secs)	
	BCDG(L_Q|Sp|Sen)	ANS	BCDG	ANS	BCDG	ANS
500 | 2.74 | 0	1662 (2.9-2| 0.99| 0.72)	46	-8.18195357 2	2.42-2	5.5	11.5
1000 | 4.15 | 0	8701 (1.5-2| 0.99| 0.99)	87	-4.32170724 2	8.60-5	145.7	117.7
1500 | 4.63 | 0	12661 (1.4-2| 0.99| 0.98)	84	-4.35887269 2	1.65-3	491.9	371.1
2000 | 5.14 | 0	18781 (1.2-2| 0.98| 0.97)	93	-2.80176287 2	6.03-4	1285.1	953.9
500 | 1.97 | 60702	3601 (2.9-2| 1.00| 0.76)	619	-8.42619444 2	-1.54-1	12.5	146.9
1000 | 3.33 | 241887	11341 (1.6-2| 1.00| 0.99)	807	-4.45131714 2	-3.53-3	195.8	1053.4
1500 | 3.71 | 542496	13321 (1.4-2| 1.00| 0.99)	969	-4.63013088 2	-1.21-1	528.2	3839.5
2000 | 4.13 | 961274	20681 (1.3-2| 1.00| 0.99)	1256	-3.19691367 2	-7.90-2	1448.8	10845.3

Conclusions & Future Work

1. The BCGD method may be viewed as a hybrid of gradient-projection and SOR methods, or as a block-coordinate version of descent methods.

2. The method achieves linear convergence, and terminates in $O(n^5/\epsilon)$ operations with an ϵ-optimal solution.

3. Preliminary numerical experience suggests that our method is efficient to solve the dual formulation of large-scale covariance selection problems especially with a lot of constraints.

4. Can the complexity bound $O(n^5/\epsilon)$ be sharpened?

5. Are there other efficient choices of $\mathcal{J}^k$ ensuring the convergence?

Thank you!

Yun S., Tseng, P., and Toh K.-C., A block coordinate gradient descent method for regularized convex separable optimization and covariance selection.