# A Block Coordinate Gradient Descent Method for Log-determinant Semidefinite Programming

Sangwoon Yun

Computational Sciences
Korea Institute for Advanced Study

October 23, 2010
2010 Global KMS International Conference
(Joint work with Paul Tseng (UW) and Kim-Chuan Toh (NUS))

# Outline

- Sparse Covariance Selection
- Block Coordinate Gradient Descent Method
- Convergence Results
- Numerical Experience
- Latent Variable Graphical Model Selection
- Conclusions & Future Work

# Sparse 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 $\Sigma$ is nonsingular, to estimate $\Sigma$ from $m$ samples, consider log-likelihood

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

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

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

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 maximum likelihood est. of $\Sigma^{-1}$.

- Covariance selection problem was first introduced by Dempster (1972). Since then, cov. sel. model has become a common statistical tool to distinguish direct from indirect interactions among a set of variables.
- Gaussian Graphical Model (GGM, Laurizen '96, Edwards '00) is the graphical interpretation of covariance selection model.
  In this GGM, the inverse of the covariance matrix is assumed to be sparse, and the sparsity pattern reveals the conditional independence relations satisfied by the variables.
- In the research of dependency networks of genome data (sparse gene association network exhibited in GGM can help to explain 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, sample covariance matrix $\hat{\Sigma}$ is singular.

Sparse 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 $\ell_1$-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:

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

- If conditional independence structure between all the variables are given,
  - 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:

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

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:

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

- The dual problem can be expressed:

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

  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 $n \leq 100$)
- 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

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

(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 Form

$$\min_{X \in \mathcal{S}^n} \quad F(X) := f(X) + P(X),$$

where

- for primal problem,

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

- for dual problem,

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

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

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

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 $\mathcal{H}(D) = (H_{ij}D_{ij})_{ij}$ 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 $P_{ij}$ 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 $\alpha$ to be the largest element of $\{\beta^k\}_{k=0,1,\ldots}$ satisfying

$$F(X+\alpha D) \le F(X) + \alpha\sigma q_{\mathcal{H}}(X;\mathcal{J}) \quad (0<\beta<1, 0<\sigma<1).$$

The limited minimization rule

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

where $0 < s < \infty$, can also be used (especially for dual).

## Choose $\mathcal{J}$:

- 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.

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

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

(ex: $\mathcal{J}^k = \{(i,j), (j,i) \mid i = 1, ..., n\}$ where $j = k + 1 (\mathrm{mod}\ n)$)

- Update only one column (and corresponding row) 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

- $0 < \underline{\lambda} \leq \lambda_{\text{min}}(\mathcal{H})$ and $\lambda_{\text{max}}(\mathcal{H}) \leq \bar{\lambda}$, and $\alpha$ 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 $F$.

**Iteration Complexity**

- For the dual problems, BCGD method can be implemented to achieve $\epsilon$-optimality in
$$O\left(\frac{n^5}{\epsilon}\right)$$
operations.
- Worst-case arithmetic cost of the first-order method proposed by Lu to achieve $\epsilon$-optimality for unconstrained problem is $O(n^4/\sqrt{\epsilon})$ operations.

# Numerical Experience

The dual form:

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

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

- 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 $10^{-10} \le ((X^k)^{-1})_{jj} \le 10^{10}$ 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 $\alpha^k$ by the limited minimization rule.

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

  - 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 $w^k,\ r^k \in \Re$.
  - Quantity $-\log\det(X^k+\alpha D^k)(+\log\det(X^k))$ is minimized when

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

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

- Termination Criterion:

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

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

## Generating test problems:

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

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

- Finally, obtain randomly generated sample covariance matrix:

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

where $B = \Sigma + 0.15 \frac{\|\Sigma\|_F}{\|\Xi\|_F} \Xi$ and $\Xi \in \mathcal{S}^n$ is a random matrix whose elements are drawn from the uniform distribution on the interval $[-1, 1]$.

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

# 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 $5 \times 10^{-2}$ in absolute value to 0.

| $n$ \| density(%) \| **card**$(V)$ | iteration count | | primal objective value | | time (secs) | |
|---|---|---|---|---|---|---|
| | BCDG$(L_Q\|\mathrm{Sp}\|\mathrm{Sen})$ | 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 |

$\mathrm{L}_Q := \frac{1}{n}\|\Sigma X - I\|_F$ (quality of the approximation of $\Sigma^{-1}$ by $X$)
$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$ and $\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ (quality of sparsity pattern)

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 $\Sigma^{-1}$.

# Latent Variable Graphical Model Selection

- In many applications throughout science and engineering, a challenge is that some of the relevant variables may be hidden or unobserved.
- Assume that we have samples of a subset of a collection of random variables. No information is provided about the number of latent variables, nor the relationship between the latent & observed variables.
- Latent variables pose a significant difficulty for model selection.

The proposed optimization problem:

$$\min_{S,L\in\mathcal{S}^n, S-L\succ 0, L\succeq 0} -\log\det(S-L)+\left\langle\hat{\Sigma},\, S-L\right\rangle+\lambda\left[\sum_{i,j=1}^{n}\rho_{ij}|S_{ij}|+\|L\|_*\right].$$

This modeling framework consistently estimates both the number of hidden components (low rank) and the conditional graphical model structure (sparsity pattern) among the observed variables.

# 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 $\epsilon$-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 BCGD method simply be extended to solve the problem arising in latent variable graphical model selection?

5. Other efficient algorithms for the problem arising in latent variable graphical model selection?

Thank you!

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