OR 6327: Semidefinite Programming. Spring 2012.

Comments on the final exam.

1. a) Since $A \succ 0$, it has a Cholesky factorization $A = LL^T$, with L invertible. Then by the change of variable $W := L^T X L$, the primal problem becomes $\min\{D \bullet W : I \bullet W = \beta, W \succeq 0\}$, where $D := L^{-1} C L^{-T}$. Next, D is symmetric, so it has an eigenvalue decomposition $D = Q \Delta Q^T$, with $\Delta = \mathrm{Diag}\,(d)$. Another change of variable $Z := Q^T W Q$ leads to $\min\{\Delta \bullet Z : I \bullet Z = \beta, Z \succeq 0\}$. Since Δ and I are diagonal, without loss of generality the optimal Z is, and then it is easily seen that $Z = \beta e_i e_i^T$ is optimal, where $d_i = \min\{d_j\}$. Correspondingly, $W = \beta Q e_i e_i^T Q^T$ and $X = \beta L^{-T} Q e_i e_i^T Q^T L^{-1}$ are optimal in their respective problems.

The dual is $\max\{\beta\eta : \eta A + S = C, \eta \in \mathbb{R}, S \succeq 0\}$. Since L and Q as above are invertible, the constraint is equivalent to $\eta I + L^{-1} S L^{-T} = D$ and then $\eta I + Q^T L^{-1} S L^{-T} Q = \Delta$. Since I and Δ are diagonal, we can set $\eta = d_i$ and thus $S = LQ(\Delta - d_i I) Q^T L^T \succeq 0$.

To check, we note that both are feasible and both have the same objective value $\beta \min\{d_j\}$, or that both are feasible and $X \bullet S = (e_i e_i^T) \bullet (\Delta - d_i I) = 0$.

b) If $f(X) := X^{-1} \bullet W$, then $Df(X)[H] = -X^{-1} H X^{-1} \bullet W$, and $D^2 f(X)[H, H] = 2X^{-1} H X^{-1} H X^{-1} \bullet W$. Since X^{-1} and W are positive definite, they have positive definite square roots $X^{-1/2}$ and $W^{1/2}$, and hence this quantity is $2 \quad \mathrm{trace}\,(X^{-1} H X^{-1} H X^{-1} W) = 2 \quad \mathrm{trace}\,(X^{-1/2} H X^{-1} W^{1/2} W^{1/2} X^{-1} H X^{-1/2}) = 2\,\|W^{1/2} X^{-1} H X^{-1/2}\|_F^2 \geq 0$. Since the second derivative in any direction is nonnegative, f is convex.

2. You generally all did well on this one.

3. a) In addition to the equations maintaining feasibility, $\mathcal{A}^* \Delta y + \Delta S = 0$, $\mathcal{A} \Delta X = 0$, we have the equation $S \Delta X S + \Delta S X S + S X \Delta S - \mu \Delta S = -SXS + \mu S$ (for $\mu > 0$) or $(S \odot S) \Delta X + 2(SX \odot I) \Delta S = -SXS$ for $\mu = 0$.

b) You can pre- and post-multiply the last equation by S^{-1} if you like, or leave it as it is; in any case, it bears a strong resemblance to the HRVW-KSH equation, the difference being the factor 2.

Note that there is a unique solution to these equations by our general resul, since $\mathcal{E}$ is invertible and $\mathcal{E}^{-1} \mathcal{F}$ is positive definite.

Also, it is easy to see that $(\Delta X, 2\Delta y, 2\Delta S)$ satisfies the equations for the HRVW-KSH direction, so the new direction is obtained from that by just scaling the dual parts by 1/2.

c) If (y, S) is not feasible, then the right-hand side of the first equation above becomes $b - \mathcal{A}^* y - S \neq 0$, so $(\Delta X, 2\Delta y, 2\Delta S)$ does *not* satisfy this equation. Hence we no longer get a scaled HRVW-KSH direction.

OVER

4. a) There were several ways to do this, and you generally did well. One way is to decompose the 2×2 block matrix M as the product of three matrices, as in Fact 11 in Lecture 3 (note that there was an incorrect sign in the notes, since corrected: thanks, Mark!).

b) The key here was to interpret the covariance matrix as M^{-1} in part (a) (the use of U for the covariance matrix of the observable variables was intended to suggest this). The we expect M, and hence A, to be sparse. Also, by (a), $U^{-1} = A - BC^{-1}B^T$, the sum of a sparse matrix A and a low-rank (at most k) matrix $-BC^{-1}B^T$.

c) I wanted a little explanation of why minimizing the nuclear norm encouraged low rank, while minimizing the L_1-norm of the entries encouraged sparsity. A tighter formulation would maybe have $-L + S = D$ with both L and S positive semidefinite, but this complicates the algorithm. (For this formulation, where also a maximum-likelihood term is used because the covariance matrix is not known exactly, see Chandrasekharan, Parrilo, and Willsky, "Latent variable graphical model selection via convex optimization.") With or without this wrinkle, you really should have a tuning parameter weighting the L_1-norm; remarkably, this can be chosen to be $1/\sqrt{n}$, and exact recovery is possible under weak conditions. (See Candes, Li, Ma, and Wright, "Robust principal component analysis?")

d) For the SDP reformulation, you either add the constraint

$$\begin{bmatrix} U & L \\ L & V \end{bmatrix} \succeq 0$$

and replace $\|L\|_*$ by $(I \bullet U + I \bullet V)/2$, or (since L is symmetric) add $-W \preceq L \preceq W$ and replace it by $I \bullet W$. Also, add the constraints $-Z \le S \le Z$ and replace $|||S|||_1$ by $ee^T \bullet Z$.

These formulations have a large number of constraints and variables, so a good algorithm is the alternating direction augmented Lagrangian method, combining ideas of Lectures 27 and 28. Introduce the function

$$g(L, S, Y) := \|L\|_* + |||S|||_1 - Y \bullet (L + S - D) + \frac{\beta}{2}\|L + S - D\|_F^2.$$

Then the algorithm is to choose L_0, S_0, and Y_0 suitably, and at iteration k, set $L_{k+1} := \arg\min_L g(L, S_k, Y_k)$, then $S_{k+1} := \arg\min_S g(L_{k+1}, S, Y_k)$, and finally $Y_{k+1} := Y_k - \beta(L_{k+1} + S_{k+1} - D)$.

I wanted a little explanation of how the updates could be carried out, as in Lectures 27 and 28. (Also, while I didn't mention it in Lecture 27, note that the update of Y is chosen so that, if g had been exactly minimized as a function of L and S, the new Y would be such that the gradient of the ordinary Lagrangian with this new Y with respect to L and S would be zero. It is also a steepest ascent step with suitable step size for the dual function.)

Have a great summer!