16th IFAC Symposium on System Identification
The International Federation of Automatic Control
Brussels, Belgium. July 11-13, 2012

Convex optimization techniques in system identification

Lieven Vandenberghe *

* *Electrical Engineering Department, UCLA, Los Angeles, CA 90095*
(Tel: 310-206-1259; e-mail: vandenbe@ee.ucla.edu)

Abstract: In recent years there has been growing interest in convex optimization techniques for system identification and time series modeling. This interest is motivated by the success of convex methods for sparse optimization and rank minimization in signal processing, statistics, and machine learning, and by the development of new classes of algorithms for large-scale nondifferentiable convex optimization.

1. INTRODUCTION

Low-dimensional model structure in identification problems is typically expressed in terms of matrix rank or sparsity of parameters. In optimization formulations this generally leads to non-convex constraints or objective functions. However, formulations based on convex penalties that indirectly minimize rank or maximize sparsity are often quite effective as heuristics, relaxations, or, in rare cases, exact reformulations. The best known example is 1-norm regularization in sparse optimization, *i.e.*, the use of the 1-norm $\|x\|_1$ in an optimization problem as a substitute for the cardinality (number of nonzero elements) of a vector x. This idea has a rich history in statistics, image and signal processing [Rudin et al., 1992, Tibshirani, 1996, Chen et al., 1999, Efron et al., 2004, Candès and Tao, 2007], and an extensive mathematical theory has been developed to explain when and why it works well [Donoho and Huo, 2001, Donoho and Tanner, 2005, Candès et al., 2006b, Candès and Tao, 2005, Candès et al., 2006a, Candès and Tao, 2006, Donoho, 2006, Tropp, 2006]. Several excellent surveys and tutorials on this topic are available; see for example [Romberg, 2008, Candès and Wakin, 2008, Elad, 2010].

The 1-norm used in sparse optimization has a natural counterpart in the nuclear norm for matrix rank minimization. Here one uses the penalty function $\|X\|_*$ where $\|\cdot\|_*$ denotes the nuclear norm (sum of singular values) as a substitute for $\mathbf{rank}(X)$. Applications of nuclear norm methods in system theory and control were first explored by [Fazel, 2002, Fazel et al., 2004], and have recently gained in popularity in the wake of the success of 1-norm techniques for sparse optimization [Recht et al., 2010]. Much of the recent work in this area has focused on the low-rank matrix completion problem [Candès and Recht, 2009, Candès and Plan, 2010, Candès and Tao, 2010, Mazumder et al., 2010], *i.e.*, the problem of identifying a low-rank matrix from a subset of its entries. This problem has applications in collaborative prediction [Srebro et al., 2005] and multi-task learning [Pong et al., 2011]. Applications of nuclear norm methods in system identification are discussed in [Liu and Vandenberghe, 2009a, Grossmann et al., 2009, Mohan and Fazel, 2010, Gebraad et al., 2011, Fazel et al., 2011].

The 1-norm and nuclear norm techniques can be extended in several interesting ways. The two types of penalties can be combined to promote sparse-plus-low-rank structure in matrices [Candès et al., 2011, Chandrasekaran et al., 2011]. Structured sparsity, such as group sparsity or hierarchical sparsity, can be induced by extensions of the 1-norm penalty [Bach et al., 2012, Jenatton et al., 2011, Bach et al., 2011]. Finally, Chandrasekaran et al. [2010] and Bach [2010] describe systematic approaches for constructing convex penalties for different types of nonconvex structural constraints.

In this tutorial paper we discuss a few applications of convex methods for structured rank minimization and sparse optimization, in combination with classical ideas from system identification and signal processing. We focus on subspace algorithms for system identification and topology selection problems in graphical models. The second part of the paper (section 4) provides a short survey of available convex optimization algorithms.

2. SYSTEM IDENTIFICATION

Subspace methods in system identification and signal processing rely on singular value decompositions (SVDs) to make low-rank matrix approximations [Ljung, 1999]. The structure in the approximated matrices (for example, Hankel structure) is therefore lost during the low-rank approximation. A convex optimization formulation based on the nuclear norm penalty offers an interesting alternative, because it promotes low rank while preserving linear matrix structure. An additional benefit of an optimization formulation is the possibility of adding other convex regularization terms or constraints on the optimization variables.

As an illustration, consider the input-output equation used as starting point in many subspace identification methods:

$$Y = \mathcal{O}X + HU.$$

The matrices U and Y are block Hankel matrices constructed from a sequence of inputs $u(t)$ and outputs $y(t)$ of a state space model

$$x(t+1) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t),$$

978-3-902823-06-9/12/$20.00 © 2012 IFAC

10.3182/20120711-3-BE-2027.00244

and the columns of X form a sequence of states $x(t)$. The matrix H depends on the system matrices, and $\mathcal{O}$ is an extended observability matrix [Verhaegen and Verdult, 2007, p.295] A simple subspace method consists in forming the Hankel matrices U and Y and then projecting the rows of Y on the nullspace of U. If the data are exact and a persistence of excitation assumption holds, the rank of the projected output matrix is equal to the system order and from it a system realization is easily computed. When the input-output data are not exact, one can use a singular value decomposition of the projected output Hankel matrix to estimate the order and compute a system realization. However, as mentioned, this step destroys the Hankel structure in Y and U. The nuclear norm penalty on the other hand can be used as a convex heuristic for indirectly reducing the rank, while preserving linear structure. For example, if the inputs are exactly known and the measured outputs $y_{\rm m}(t)$ are subject to error, one can solve the convex problem

$$\text{minimize } \|YQ\|_* + \rho \sum_t \|y(t) - y_{\rm m}(t)\|_2^2$$

where the columns of Q form a basis of the nullspace of U and ρ is a positive weight. The optimization variables are the model outputs $y(t)$ and the matrix Y is a Hankel matrix constructed from the model outputs $y(t)$. This is a convex optimization problem that can be solved via semidefinite programming. We refer the reader to [Liu and Vandenberghe, 2009a,b] for more details and numerical results. As an important advantage, the optimization formulation can be extended to include convex contraints on the model outputs. Another promising application is identification with missing data [Ding et al., 2007, Grossmann et al., 2009].

3. GRAPHICAL MODELS

In a graphical model of a normal distribution $x \sim \mathcal{N}(0, \Sigma)$ the edges in the graph represent the conditional dependence relations between the components of x. The vertices in the graph correspond to the components of x; the absence of an edge between vertices i and j indicates that x_i and x_j are independent, conditional on the other entries of x. Equivalently, vertices i and j are connected if there is a nonzero in the i, j position of the inverse covariance matrix Σ^{-1}.

A key problem in the estimation of the graphical model is the selection of the topology. Several authors have addressed this problem by adding a 1-norm penalty to the maximum likelihood estimation problem, and solving

$$\text{minimize } \mathbf{tr}\, CX - \log\det X + \rho\|X\|_1. \tag{1}$$

Here X denotes the inverse covariance Σ^{-1}, the matrix C is the sample covariance matrix, and $\|X\|_1 = \sum_{ij} |X_{ij}|$. See [Meinshausen and Bühlmann, 2006, Banerjee et al., 2008, Ravikumar et al., 2008, Friedman et al., 2008, Lu, 2009, Scheinberg and Rish, 2009, Yuan and Lin, 2007, Duchi et al., 2008, Li and Toh, 2010, Scheinberg and Ma, 2012].

Graphical models of the conditional independence relations can be extended to Gaussian vector time series [Brillinger, 1996, Dahlhaus, 2000]. In this extension the topology of the graph is determined by the sparsity pattern of the inverse spectral density matrix

$$S(\omega) = \sum_{k=-\infty}^{\infty} R_k e^{\mathrm{j}k\omega},$$

with $R_k = \mathbf{E}\, x(t+k)x(t)^T$. Using this characterization, one can formulate extensions of the regularized maximum likelihood problem (1) to vector time series. In [Songsiri et al., 2010, Songsiri and Vandenberghe, 2010] autoregressive models

$$x(t) = -\sum_{k=1}^{p} A_k x(t-k) + v(t), \qquad v(t) \sim \mathcal{N}(0, \Sigma),$$

were considered, and convex formulations were presented for the problem of estimating the parameters A_k, Σ, subject to conditional independence constraints, and of estimating the topology via a 1-norm type regularization. The topology selection problem leads to the following extension of (1):

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(CX) - \log\det X_{00} + \rho h(X) \\ \text{subject to} & X \succeq 0. \end{array} \tag{2}$$

The variable X is a $(p+1) \times (p+1)$ block matrix with blocks of size $n \times n$ (the length of the vector $x(t)$), and X_{00} is the leading block of X. The penalty h is chosen to encourage a common, symmetric sparsity pattern for the diagonal sums

$$\sum_{i=0}^{p-k} X_{i,i+k}, \quad k = 0, 1, \ldots, p,$$

of the blocks in X.

An extension to ARMA processes is studied by Avventiy et al. [2010].

4. ALGORITHMS

For small and medium sized problems the applications discussed in the previous sections can be handled by general-purpose convex optimization solvers, such as the modeling packages CVX [Grant and Boyd, 2007] and YALMIP [Löfberg, 2004], and general-purpose conic optimization packages. In this section we discuss algorithmic approaches that are of interest for large problems that fall outside the scope of the general-purpose solvers.

4.1 Interior-point algorithms

Interior-point algorithms are known to attain a high accuracy in a small number of iterations, fairly independent of problem data and dimensions. The main drawback is the high linear algebra complexity per iteration associated with solving the Newton equations that determine search directions. However sometimes problem structure can be exploited to devise dedicated interior-point implementations that are significantly more efficient than general-purpose solvers.

A simple example is the 1-norm approximation problem

$$\text{minimize } \|Ax - b\|_1$$

with A of size $m \times n$. This can be formulated as a linear program (LP)

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} y_i \\ \text{subject to} & \begin{bmatrix} A & -I \\ -A & -I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} b \\ -b \end{bmatrix}, \end{array}$$

at the expense of introducing m auxiliary variables and $2m$ linear inequality constraints. By taking advantage of the structure in the inequalities, each iteration of an interior-point method for the LP can be reduced to solving linear systems $A^T D A \Delta x = r$ where D is a positive diagonal matrix. As a result, the complexity of solving the 1-norm approximation problem using a custom interior-point solver is roughly the equivalent of a small number of weighted least-squares problems.

A similar result holds for the nuclear norm approximation problem

$$\text{minimize } \|A(x) - B\|_* \tag{3}$$

where $A(x)$ is a matrix valued function of size $p \times q$ and x is an n-vector of variables. This problem can be formulated as a semidefinite program (SDP)

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, U + \mathbf{tr}\, V \\ \text{subject to} & \begin{bmatrix} U & (A(x) - B)^T \\ A(x) - B & V \end{bmatrix} \succeq 0 \end{array} \tag{4}$$

with variables x, U, V. The very larger number of variables ($O(p^2)$ if we assume $p \geq q$) makes the nuclear norm optimization problem very expensive to solve by general-purpose SDP solvers. A specialized interior-point solver for the SDP is described in [Liu and Vandenberghe, 2009a], with a linear algebra cost per iteration of $O(n^2pq)$ if $n \geq \max\{p, q\}$. This is comparable to solving the matrix approximation problem in Frobenius norm, *i.e.*, minimizing $\|A(x) - B\|_F$, and the improvement makes it possible to solve nuclear norm problems with p and q on the order of several hundred by an interior-point method.

We refer the reader to the book chapter [Andersen et al., 2012] for additional examples of special-purpose interior-point algorithms.

4.2 Nonlinear optimization methods

Burer and Monteiro Burer and Monteiro [2003, 2005] have developed a large-scale method for semidefinite programming, based on substituting a low-rank factorization for the matrix variable and solving the resulting nonconvex problem by an augmented Lagrangian method. Adapted to the SDP (4), the method amounts to reformulating the problem as

$$\begin{array}{ll} \text{minimize} & \|L\|_F^2 + \|R\|_F^2 \\ \text{subject to} & A(x) - B = LR^T \end{array} \tag{5}$$

with variables x, $L \in \mathbf{R}^{p \times r}$, $R \in \mathbf{R}^{q \times r}$, where r is a upper bound on the rank of $A(x) - b$ at optimum. Recht et al. [2010] discuss in detail Burer and Monteiro's method in the context of nuclear norm optimization.

4.3 Proximal gradient algorithms

The proximal gradient algorithm is an extension of the gradient algorithm to problems with simple constraints or with simple nondifferentiable terms in the cost function. It is less general than the subgradient algorithm, but it is typically much faster and it handles many types of nondifferentiable problems that occur in practice.

The proximal gradient algorithm applies to a convex problem of the form

$$\text{minimize } f(x) = g(x) + h(x), \tag{6}$$

in which the cost function f is split in two components g and h, with g differentiable and h a 'simple' nondifferentiable function. 'Simple' here means that the *prox-operator* of h, defined as the mapping

$$\mathbf{prox}_{th}(x) = \underset{u}{\text{argmin}} \left(h(u) + \frac{1}{2t}\|u - x\|_2^2 \right)$$

(with $t > 0$), is inexpensive to compute. It can be shown that if h is closed and convex, then $\mathbf{prox}_{th}(x)$ exists and is unique for every x.

A typical example is $h(x) = \|x\|_1$. Its prox-operator is the element-wise 'soft-thresholding'

$$\mathbf{prox}_{th}(x)_i = \begin{cases} x_i - t & \text{if } x_i \geq t \\ 0 & \text{if } -t \leq x_i \leq t \\ x_i + t & \text{if } x_i \leq -t. \end{cases}$$

Constrained optimization problems

$$\begin{array}{ll} \text{minimize} & g(x) \\ \text{subject to} & x \in C \end{array}$$

can be brought in the form (6) by defining $h(x) = I_C(x)$, the indicator function of C (*i.e.*, $I_C(x) = 0$ if $x \in C$ and $I_C(x) = +\infty$ if $x \notin C$). The prox-operator for I_C is the Euclidean projection on C. Prox-operators share many of the properties of Euclidean projections on closed convex sets. For example, they are nonexpansive, *i.e.*,

$$\|\mathbf{prox}_{th}(x) - \mathbf{prox}_{th}(y)\|_2 \leq \|x - y\|_2$$

for all x, y. (See Moreau [1965].)

The proximal gradient method for minimizing (6) uses the iteration

$$x^+ = \mathbf{prox}_{th}\left(x - t\nabla g(x)\right)$$

where $t > 0$ is a step size. The proximal gradient update consists of a standard gradient step for the differentiable term g, followed by an application of the prox-operator associated with the non-differentiable term h. It can be motivated by noting that x^+ is the minimizer of the function

$$h(y) + g(x) + \nabla g(x)^T(y - x) + \frac{1}{2t}\|y - x\|_2^2$$

over y, so x^+ minimizes an approximation of f, obtained by adding to h a simple local quadratic model of g.

It can be shown that if ∇g is Lipschitz continuous with constant L, then the suboptimality $f(x^{(k)}) - f^\star$ decreases to zero as $O(1/k)$ [Nesterov, 2004, Beck and Teboulle, 2009]. Recently, faster variants of the proximal gradient

method with an $1/k^2$ rate convergence, under the same assumptions and with the same complexity per step, have been developed [Nesterov, 2004, 2005, Beck and Teboulle, 2009, Tseng, 2008, Becker et al., 2011].

The (accelerated) proximal gradient methods are well suited for problems of the form

$$\text{minimize } g(x) + \|x\|$$

where g is differentiable with a Lipschitz-continuous gradient. Most common norms have easily computed prox-operators, and the following property is useful when computing the prox-operator of a norm $h(x) = \|x\|$:

$$\mathbf{prox}_{th}(x) = x - tP_B(x/t),$$

where P_B is Euclidean projection on the unit ball in the dual norm.

In other applications it is advantageous to apply the proximal gradient method to the dual problem. Consider for example an optimization problem

$$\text{minimize } f(x) + \|Ax - b\|$$

with f strongly convex. Reformulating this problem as

$$\begin{array}{ll} \text{minimize} & f(x) + \|y\| \\ \text{subject to} & y = Ax - b \end{array} \tag{7}$$

and taking the Lagrange dual, gives

$$\begin{array}{ll} \text{maximize} & b^T z - f^*(A^T z) \\ \text{subject to} & \|z\|_{\mathrm{d}} \leq 1 \end{array}$$

where $f^*(u) = \sup_x(u^T x - f(x))$ is the conjugate of f and $\|\cdot\|_{\mathrm{d}}$ is the dual norm of $\|\cdot\|$. It can be shown that if f is strongly convex, then f^* is differentiable with a Lipschitz continuous gradient. If projection on the unit ball of the dual norm is inexpensive, the dual problem is therefore readily solved by a fast gradient projection method.

An extensive library of fast proximal-type algorithms is available in the MATLAB software package TFOCS [Becker et al., 2010].

4.4 ADMM

The *Alternating Direction Method of Multipliers* (ADMM) was proposed in the 1970s as a simplified version of the augmented Lagrangian method. It is a simple and often very effective method for large-scale or distributed optimization, and has recently been applied successfully to the regularized covariance selection problem mentioned above [Scheinberg et al., 2010, Scheinberg and Ma, 2012]. The recent survey by Boyd et al. [2011] gives an overview of the theory and applications of ADMM. Here we limit ourselves to a description of the method when applied to a problem of the form (7). The ADMM iteration consists of two alternating minimization steps (over x and y) of the augmented Lagrangian

$$L(x, y, z) = f(x) + \|y\| + z^T(y - Ax + b) + \frac{t}{2}\|y - Ax + b\|_2^2,$$

followed by an update

$$z := z + t(y - Ax - b)$$

of the dual variable z. The complexity of minimizing over x depends on the properties of f. If f is quadratic, for example, it reduces to a least-squares problem. The minimization of the augmented Lagrangian over y reduces to the evaluation of the prox-operator of the norm $\|\cdot\|$.

A numerical comparison of the ADMM and proximal gradient algorithms for nuclear norm minimization can be found in the recent paper by Fazel et al. [2011].

5. SUMMARY

Advances in algorithms for large-scale nondifferentiable convex optimization are leading to a greater role of convex optimization in system identification and time series modeling. These techniques are based on formulations that incorporate convex penalty functions that promote low-dimensional model structure (such as sparsity or rank). Similar techniques have been used extensively in signal processing, image processing, and machine learning. While at this point theoretical results that characterize the success of these convex heuristics in system identification are limited, the extensive theory that supports 1-norm techniques in sparse optimization, gives hope that progress can be made in our understanding of similar techniques for system identification as well.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grants No. ECCS-0824003 and ECCS-1128817. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

M. S. Andersen, J. Dahl, Z. Liu, and L. Vandenberghe. Interior-point methods for large-scale cone programming. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 55–83. MIT Press, 2012.

E. Avventiy, A. Lindquist, and B. Wahlberg. Graphical models of autoregressive moving-average processes. In *The 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010)*, July 2010.

F. Bach. Structued sparsity-inducing norms through submodular functions. 2010. Available from `arxiv.org/abs/1008.4220`.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 19–53. MIT Press, 2012.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

S. Becker, E. J. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. 2010. `arxiv.org/abs/1009.2065`.

S. Becker, J. Bobin, and E. Candès. NESTA: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of mulitipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

D. R. Brillinger. Remarks concerning graphical models for time series and point processes. *Revista de Econometria*, 16:1–23, 1996.

S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming (Series B)*, 95 (2), 2003.

S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming (Series A)*, 103(3), 2005.

E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005.

E. J. Candès and T. Tao. Near-optimal signal recovery from random projections and universal encoding strategies. *IEEE Transaction on Information Theory*, 52(12), 2006.

E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006a.

E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59 (8):1207–1223, 2006b.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58 (3), 2011.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. 2010. arXiv:1012.0621v1.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

R. Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172, 2000.

T. Ding, M. Sznaier, and O. Camps. A rank minimization approach to fast dynamic event detection and track matching in video sequences. In *Proceedings of the 46th IEEE conference on decision and control*, 2007.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined systems by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse Gaussians. In *Proceedings of the Conference on Uncertainty in AI*, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proceedings of American Control Conference*, pages 3273–3278, 2004.

M. Fazel, T. K. Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applictions to system identification and realization. 2011. Submitted.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

P. M. O. Gebraad, J. W. van Wingerden, G. J. van der Veen, and M. Verhaegen. LPV subspace identification using a novel nuclear norm regularization method. In *Proceedings of the American Control Conference*, pages 165–170, 2011.

M. Grant and S. Boyd. *CVX: Matlab software for disciplined convex programming (web page and software)*. `http://stanford.edu/~boyd/cvx`, 2007.

C. Grossmann, C. N. Jones, and M. Morari. System identification via nuclear norm regularization for simulated bed processes from incomplete data sets. In *Proceedings of the 48th IEEE Conference on Decision and Control*, pages 4692–4697, 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

L. Li and K.-C. Toh. An inexact interior point method for L1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2:291–315, 2010.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31:1235–1256, 2009a.

Z. Liu and L. Vandenberghe. Semidefinite programming methods for system realization and identification. In *Proceedings of the 48th IEEE Conference on Decision and Control*, pages 4676–4681, 2009b.

L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, New Jersey, second edition, 1999.

J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11: 2287–2322, 2010.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In *Proceedings of the American Control Conference (ACC)*, pages 2953–2959, 2010.

J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Math. Soc. France*, 93:273–299, 1965.

Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming Series A*, 103:127–152, 2005.

T. K. Pong, P. Tseng, Shuiwang Ji, and J. Ye. Trace norm regularization: reformulations, algorithms, and multitask learning. *SIAM Journal on Optimization*, 20(6): 3465–3489, 2011.

R. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, 2008. `arxiv.org/abs/0811.3628`.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

J. Romberg. Imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):14–20, 2008.

L. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

K. Scheinberg and S. Ma. Optimization methods for sparse inverse covariance selection. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 455–477. MIT Press, 2012.

K. Scheinberg and I. Rish. SINCO - a greedy coordinate ascent method for sparse inverse covariance selection problem. Technical report, 2009. IBM Resesarch Report.

K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2101–2109. 2010.

J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 11:2671–2705, 2010.

J. Songsiri, J. Dahl, and L. Vandenberghe. Graphical models of autoregressive processes. In Y. Eldar and D. Palomar, editors, *Convex Optimization in Signal Processing and Communications*, pages 89–116. Cambridge University Press, Cambridge, 2010.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, Cambridge, MA, 2005.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008.

M. Verhaegen and V. Verdult. *Filtering and System Identification*. Cambridge University Press, 2007.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19, 2007.