SPARSE COVARIANCE ESTIMATION UNDER KRONECKER PRODUCT STRUCTURE

Theodoros Tsiligkaridis[*] *and Alfred O. Hero III*[*,†]

University of Michigan, [*]EECS Dept. and [†] Dept. Statistics, Ann Arbor, USA
{ttsili,hero}@umich.edu

ABSTRACT

We introduce a sparse covariance estimation method for the high dimensional setting when the covariance matrix decomposes as a Kronecker product, i.e., $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, and the observations are Gaussian. We propose an ℓ_1 penalized maximum-likelihood approach to solve this problem. The dual formulation motivates an iterative algorithm (penalized flip-flop; FFP) based on a block coordinate-descent approach. Although the ℓ_1-penalized log-likelihood function (objective function) is non-convex in general and non-smooth, we show that FFP converges to a local maximum under relatively mild assumptions. For the fixed dimension case, large-sample statistical consistency is proved and a rate of convergence bound is derived. Simulations show that FFP outperforms its non-penalized counterpart and the naive Glasso algorithm for sparse Kronecker-decomposable covariance matrix.

Index Terms— high dimensional inference, penalized maximum likelihood, direct product, Glasso, dual optimization

1. INTRODUCTION

Covariance estimation is a problem of interest in many different disciplines, including machine learning, signal processing, economics and bioinformatics. Consider a separable covariance matrix model for the observable data:

$$\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0 \tag{1}$$

where $\mathbf{A}_0$ is a $p \times p$ positive definite matrix and $\mathbf{B}_0$ is an $q \times q$ positive definite matrix. Note that this implies that $\mathbf{\Sigma}_0$ is positive definite. Let $\mathbf{\Theta}_0 := \mathbf{\Sigma}_0^{-1}$ denote the inverse covariance, or precision matrix. As the number of parameters is reduced from $\Theta(p^2q^2)$ to $\Theta(p^2) + \Theta(q^2)$, the factorization (1) offers a significant reduction in data requirements and in complexity.

This model arises in bioinformatics applications. When trying to estimate correlations between p genes and each gene has f factors, we can regard $\mathbf{A}_0$ as the covariance matrix between genes and $\mathbf{B}_0$ as the covariance matrix between factors. Model (1) also comes up in channel modeling for wireless communications [1]. In statistics, random processes that satisfy (1) are called separable [2]. Such a covariance model arises when the observations can be written as an i.i.d process

$$\mathbf{y}[t] = \mathbf{a}[t] \otimes \mathbf{b}[t], t = 1, \ldots, n$$

where $\mathbf{a}[t]$ and $\mathbf{b}[t]$ are two zero-mean, mutually-uncorrelated random processes with covariance matrices $\mathbf{A}_0 \in \mathbb{R}^{p \times p}$ and $\mathbf{B}_0 \in \mathbb{R}^{q \times q}$.

Estimation of the Kronecker components $(\mathbf{A}_0, \mathbf{B}_0)$ often yields superior performance to estimating $\mathbf{\Sigma}_0 \in \mathbb{R}^{pq \times pq}$ itself [3]. The maximum-likelihood (ML) estimator of the Kronecker components has been studied in [4]. While the ML estimator has no known closed-form solution, the solution can be iteratively computed via the flip-flop (FF) algorithm [4, 3].

ML estimation for the situation where the Kronecker component matrices are themselves sparse has not been studied. In addition to exploiting the Kronecker factorization, we exploit sparsity by proposing an ℓ_1-penalized likelihood estimator for the Kronecker product matrix. This formulation naturally leads to an iterative algorithm, called the penalized flip-flop (FFP) algorithm, that optimizes the ℓ_1-penalized log-likelihood function in a block-coordinate manner. The first contribution of this paper is to show that the FFP algorithm converges to a local maximum of the penalized likelihood function. We believe that due to the alternating nature of the algorithm and the lack of joint convexity with respect to the block parameters, this is the most that can be proved. The second contribution of this paper is to establish statistical convergence rates of the FFP algorithm, i.e., the rate of convergence of estimator mean squared-error.

2. NOTATION

For a square matrix $\mathbf{M}$, let $|\mathbf{M}|_1 := \|\text{vec}(\mathbf{M})\|_1$ and $|\mathbf{M}|_\infty = \|\text{vec}(\mathbf{M})\|_\infty$, where $\text{vec}(\mathbf{M})$ denotes the vectorized form of $\mathbf{M}$ (columns stacked on top of each other). Define $\mathbf{M}_{i,j}$ as the (i,j)th element of $\mathbf{M}$. Define the $pq \times pq$ permutation operator $\mathbf{P}_{p,q}$ such that $\mathbf{P}_{p,q}\text{vec}(\mathbf{N}) = \text{vec}(\mathbf{N}^T)$ for any $p \times q$ matrices $\mathbf{N}$. Define the set S_{++}^p of symmetric positive definite (p.d.) $p \times p$ matrices. It can be shown that S_{++}^p is a convex set, but it is not closed [5]. Let $I(A) \in \{0, 1\}$ be the indicator of the truth of statement A.

The research reported here was partially supported by a grant from ARO, W911NF-11-1-0391, and a Rackham Merit fellowship to the first author.

3. GRAPHICAL LASSO FRAMEWORK

Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{x}[t]\}_{t=1}^{n}$, where $\mathbf{x}[t] \in \mathbb{R}^{pq}$, having mean $\mathbf{0}$ and covariance $\mathbf{\Sigma} = \mathbf{A}_0 \otimes \mathbf{B}_0$. Then, the log-likelihood is proportional to:

$$l(\mathbf{\Sigma}) := \log\det(\mathbf{\Sigma}^{-1}) - \text{tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{S}}_n), \tag{2}$$

where $\mathbf{\Sigma}$ is the positive definite covariance matrix and $\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}[t]\mathbf{x}[t]^T$ is the sample covariance matrix. Recent work [6, 7] has considered ℓ_1-penalized maximum likelihood estimators over the unrestricted class of positive definite matrices. These estimators are known as graphical lasso and the ℓ_1 penalty induces sparsity on the solution $\hat{\mathbf{\Sigma}}_n$ by solving:

$$\hat{\mathbf{\Sigma}}_n \in \arg\max_{\mathbf{\Sigma}\in S_{++}^{p}} \{l(\mathbf{\Sigma}) - \lambda|\mathbf{\Sigma}^{-1}|_1\}, \tag{3}$$

where $\lambda \geq 0$ is a regularization parameter. If $\lambda > 0$ and $\hat{\mathbf{S}}_n$ is positive definite, then $\hat{\mathbf{\Sigma}}_n$ in (3) is the unique minimizer.

For covariance matrices $\mathbf{\Sigma}$ of the form $\mathbf{A} \otimes \mathbf{B}$ (where $\mathbf{A} \in S_{++}^{p}$ and $\mathbf{B} \in S_{++}^{q}$), we have $\mathbf{\Sigma}^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ and $|\mathbf{\Sigma}^{-1}|_1 = |\mathbf{A}^{-1}|_1|\mathbf{B}^{-1}|_1$ [8]. Then, the optimization problem (3) becomes $\min_{\mathbf{X}\in S_{++}^{p},\mathbf{Y}\in S_{++}^{q}} J(\mathbf{X},\mathbf{Y})$, where

$$\begin{aligned} J(\mathbf{X},\mathbf{Y}) := \text{tr}((\mathbf{X}\otimes\mathbf{Y})\hat{\mathbf{S}}_n) - f\log\det(\mathbf{X}) \\ - p\log\det(\mathbf{Y}) + \lambda|\mathbf{X}|_1|\mathbf{Y}|_1, \end{aligned} \tag{4}$$

and we defined $\mathbf{X} = \mathbf{A}^{-1}$ and $\mathbf{Y} = \mathbf{B}^{-1}$.

Lemma 1. *Assume $\lambda \geq 0$, $\mathbf{X} \in S_{++}^{p}$ and $\mathbf{Y} \in S_{++}^{q}$. When one argument of $J(\mathbf{X},\mathbf{Y})$ is fixed, the objective function (4) is convex in the other argument* [1].

Proof. For a proof, see [9]. □

4. DUAL FORMULATION

Motivated by Lemma 1, we fix one matrix in (4) and consider the dual problem that arises as we optimize (4) over the other matrix.

Lemma 2. *Assume $\hat{\mathbf{S}}_n$ is positive definite.*

1. *Consider $J(\mathbf{X},\mathbf{Y})$ in (4) with matrix $\mathbf{X} \in S_{++}^{p}$ fixed. The dual problem for minimizing $J(\mathbf{X},\mathbf{Y})$ over $\mathbf{Y}$ is:*

$$\max_{\{\mathbf{W}:|\mathbf{W}-\frac{1}{p}\sum_{i,j=1}^{p}\mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i)|_\infty\leq\frac{\lambda|\mathbf{X}|_1}{p}\}} \log\det(\mathbf{W}). \tag{5}$$

Consider (4) with matrix $\mathbf{Y} \in S_{++}^{q}$ fixed. The dual problem for minimizing $J(\mathbf{X},\mathbf{Y})$ over X is:

$$\max_{\{\mathbf{Z}:|\mathbf{Z}-\frac{1}{q}\sum_{k,l=1}^{q}\mathbf{Y}_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)|_\infty\leq\frac{\lambda|\mathbf{Y}|_1}{q}\}} \log\det(\mathbf{Z}), \tag{6}$$

where $\overline{\hat{\mathbf{S}}_n} := \mathbf{P}_{p,q}^{T}\hat{\mathbf{S}}_n\mathbf{P}_{p,q}$.

[1] Function (4) is *not* jointly convex in $(\mathbf{X},\mathbf{Y})$, for general matrices $\mathbf{X}$ and $\mathbf{Y}$.

2. *Strong duality holds for (5) and (6).*

Proof. The proof is based on the saddle-point formulation of Lagrangian duality and is included in [9]. □

Note that both dual problems (5) and (6) have a unique solution and the maximum is attained in each one. This follows from the fact that we are maximizing a strictly concave function over a closed convex set. Lemma 2 leads to a similar result as obtained in [6], but with $\left(\frac{1}{p}\sum_{i,j=1}^{p}\mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i), \frac{\lambda|\mathbf{X}|_1}{p}\right)$ playing the role of $(\hat{\mathbf{S}}_n, \lambda)$, for the "fixed $\mathbf{X}$" subproblem (5).

5. ALGORITHM

In this section, we propose an alternating minimization algorithm based on the results obtained in Lemma 2. Since strong-duality holds for subproblems (5) and (6), we can find a δ-suboptimal solution for each subproblem by using the size of the duality gap as a stopping criterion. If $\lambda = 0$, FFP reduces to the flip-flop (FF) [3] algorithm since each Glasso step becomes superfluous.

Algorithm 1 Penalized Flip-Flop (FFP) Algorithm

1: **Input:** $\hat{\mathbf{S}}_n$, p, f, n, $\epsilon > 0$
2: **Output:** $\hat{\mathbf{\Theta}}_0$
3: Initialize $\mathbf{X}$ to be positive definite.
4: $\hat{\mathbf{\Theta}}_0 \leftarrow \mathbf{I}_{pf}$
5: **repeat**
6: $\quad\hat{\mathbf{\Theta}}_{0,\text{prev}} \leftarrow \hat{\mathbf{\Theta}}_0$
7: $\quad\mathbf{T}_X \leftarrow \frac{1}{p}\sum_{i,j=1}^{p}\mathbf{X}_{i,j}\hat{\mathbf{S}}_n(j,i)$
8: $\quad\lambda_X \leftarrow \frac{\lambda|\mathbf{X}|_1}{p}$
9: $\quad\mathbf{Y} \leftarrow \arg\min_{\mathbf{Y}\in S_{++}^{q}}\{\text{tr}(\mathbf{Y}\mathbf{T}_X) - \log\det(\mathbf{Y}) + \lambda_X|\mathbf{Y}|_1\}$
10: $\quad\mathbf{T}_Y \leftarrow \frac{1}{q}\sum_{k,l=1}^{q}\mathbf{Y}_{k,l}\overline{\hat{\mathbf{S}}_n}(l,k)$
11: $\quad\lambda_Y \leftarrow \frac{\lambda|\mathbf{Y}|_1}{q}$
12: $\quad\mathbf{X} \leftarrow \arg\min_{\mathbf{X}\in S_{++}^{p}}\{\text{tr}(\mathbf{X}\mathbf{T}_Y) - \log\det(\mathbf{X}) + \lambda_Y|\mathbf{X}|_1\}$
13: $\quad\hat{\mathbf{\Theta}}_0 \leftarrow \mathbf{X}\otimes\mathbf{Y}$
14: **until** $\|\hat{\mathbf{\Theta}}_{0,\text{prev}} - \hat{\mathbf{\Theta}}_0\| \leq \epsilon$

Steps 9 and 12 in Algorithm 1 can be solved using the Glasso algorithm of Friedman et al. [7]. Consider the "fixed $\mathbf{X}$" subproblem for concreteness. The dual program (5) is solved using Glasso resulting in a $q \times q$ matrix solution $\mathbf{W}^*$ from which $\mathbf{Y}^* = (\mathbf{W}^*)^{-1}$ can be easily obtained. This inverse exists if $\hat{\mathbf{S}}_n$ is p.d. [2]. The (scaled) duality gap at the rth inner iteration of Glasso is given by [9]:

$$\delta_{\text{gap}}^{(r)}(\mathbf{X}) = \text{tr}(\mathbf{T}_\mathbf{X}\mathbf{Y}^{(r)}) + \lambda|\mathbf{Y}^{(r)}|_1 - q$$

which is always non-negative by weak duality [5]. Once $\delta_{\text{gap}}^{(r)}(\mathbf{X}) \leq \delta$, we have a δ-suboptimal solution to the sub-

[2] All FFP iterates are p.d. if $\hat{\mathbf{S}}_n$ and the initializing matrix are p.d. [9]

problem and the iterative sub-algorithm stops. An analogous statement holds for the "fixed $\mathbf{Y}$" subproblem by symmetry.

5.1. Computational Complexity

The standard Glasso algorithm for estimating a $pq \times pq$ sparse covariance matrix has a computational complexity of order p^3q^3 for each graphical lasso procedure. On the other hand, FFP has computational complexity of order $p^3 + q^3$, which can be a significant reduction as p and q get large.

5.2. Convergence Analysis

We have established that the penalized flip-flop (FFP) algorithm converges to a local minimum under a mild assumption on the starting point. Let $J(\mathbf{X}, \mathbf{Y})$ be as defined in (4).

Theorem 1. *Assume $\hat{\mathbf{S}}_n$ is positive definite. Then, we have:*

1. *The sequence $\left\{J(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})\right\}_k$ is monotonically decreasing.*
2. *The algorithm converges to a local maximum or minimum and there are no saddle points.*
3. *As long as $(\mathbf{X}^{(0)}, \mathbf{Y}^{(0)})$ is not a local maximum, then the algorithm converges to a local minimum.*

Proof. The proof and generalizations are contained in [9]. □

6. STATISTICAL CONSISTENCY

We have also established statistical consistency and rate of convergence of the FFP algorithm for fixed dimensions p and q; specifically, $\hat{\mathbf{\Sigma}}_n \xrightarrow{p.} \mathbf{\Sigma}_0$ as $n \to \infty$, where $\hat{\mathbf{\Sigma}}_n$ denotes the penalized flip-flop (FFP) algorithm solution and $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$ denotes the true covariance matrix composed of Kronecker factors $\mathbf{A}_0$ and $\mathbf{B}_0$. Let $\mathbf{A}^{(0)}$ denote the initial guess of $\mathbf{A}_0 = \mathbf{X}_0^{-1}$. The next theorem establishes statistical consistency for the fixed dimension-large sample case with a result on rate-of-convergence.

Theorem 2. *Assume that $\mathbf{A}_0 \in S^p_{++}$ and $\mathbf{B}_0 \in S^q_{++}$. Let $\mathbf{A}^{(0)} \in S^p_{++}$ be an initial estimate of $\mathbf{A}_0$. Let $\hat{\mathbf{\Sigma}}_n$ be the estimate of $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$ generated by the FFP algorithm and assume that $\hat{\mathbf{\Sigma}}_n$ converges to the global minimum of (4). Let λ in (4) be a decreasing function of n: $\lambda_n = \frac{C_\lambda}{n^\gamma}$ for some $\gamma \geq 1/2$ and $C > 0$. Then, we have [3]:*

$$\|\hat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_0\|_F = O_p\left(\frac{\|\mathbf{\Xi}\|_2 tr(\mathbf{\Sigma}_0) + I(\gamma = \frac{1}{2})C_\lambda K_\phi}{\sqrt{n}}\right)$$

[3]Consider a sequence of real random variables X_n defined on the same probability space and a sequence of reals b_n. The notation $X_n = O_p(1)$ is defined as: $\sup_{n\in\mathbb{N}} P(|X_n| > K) \to 0$ as $K \to \infty$. The notation $X_n = O_p(b_n)$ is equivalent to $\frac{X_n}{b_n} = O_p(1)$.

where $K_\phi = K_\phi(p, f)$ is a constant independent of $\mathbf{A}^{(0)}$ and n and $\mathbf{\Xi}$ is a matrix independent of $\mathbf{A}^{(0)}$ and n (K_ϕ and $\mathbf{\Xi}$ are given explicitly in [9]).

Proof. See [9]. □

Theorem 2 gives global bounds independent of the initial condition $\mathbf{A}^{(0)}$. If one implements Algorithm 1 with a rapidly decreasing sequence of regularization parameters λ, the rate of convergence will not depend on C or K_ϕ, specifically, decaying at a rate faster than order $n^{-1/2}$. The bound implies convergence rate of at least order $n^{-1/2}$ if the regularization parameter tapers to zero fast enough. The same bound holds for $\|\mathbf{\Sigma}_n^{-1} - \mathbf{\Sigma}_0^{-1}\|_F$ for n sufficiently large [9].

7. SIMULATIONS

In this section, we present a simple Kronecker structured example to empirically compare the performance of three algorithms. The first algorithm, naive Glasso [7], simply applies the graphical lasso algorithm without imposing Kronecker structure. The second algorithm, FF, is the flip-flop algorithm [4] which iteratively computes the unpenalized ML solution. The third algorithm, FFP, is the proposed sparsity penalized flip-flop algorithm.

The true covariance matrices were held fixed as the sample size n varies. The true precision matrices $\mathbf{X}_0 := \mathbf{A}_0^{-1}$ and $\mathbf{Y}_0 := \mathbf{B}_0^{-1}$ were randomly generated p.d. matrices based on the Erdös-Rényi graph model. Figure 1 shows heatmaps of the case where both are sparse matrices. Performance assessment was based on normalized Frobenius norm error of the covariance and precision matrix estimates. The regularization parameters λ for all Glasso problems were chosen in accordance with the predictions of Thm. 2 as $c \cdot \sqrt{\frac{\log(pf)}{n}}$, where the constant c was chosen experimentally to optimize respective performances.

Figure 2 compares the root-mean squared error (rmse) performance in precision and covariance matrices as a function of n. There were a total of 10 trial runs for each n. As far as the precision matrix rmse is concerned, Glasso outperforms FF only for very small n. FFP outperforms both Glasso and FF across all n. In regards to the covariance matrix rmse, Glasso performs poorly compared to FFP and FF since it does not take into account the Kronecker product structure.

Figure 3 displays the rmse across three different sparsity scenarios, as a function of FFP iteration. Both matrices $\mathbf{X}_0$ and $\mathbf{Y}_0$ were 20×20 p.d. matrices. Comparing the last two cases, we see that although $\mathbf{Y}_0$ is dense, the sparsity of $\mathbf{X}_0$ helps a lot due to the Kronecker structure of the problem.

8. CONCLUSION

In this paper, an ℓ_1-penalized likelihood approach is proposed for estimating a sparse Kronecker-decomposable covariance

matrix given n i.i.d. Gaussian samples, leading to an iterative algorithm (FFP). Asymptotic convergence properties and large-sample statistical consistency of the FFP algorithm were established.

9. ACKNOWLEDGEMENTS

The authors thank Prof. Shuheng Zhou for helpful input.

10. REFERENCES

[1] M. Bengtsson and P. Zetterberg, "Some notes on the kronecker model," submitted for publication, April 2006.

[2] N. Lu and D. Zimmerman, "The likelihood ratio test for a separable covariance matrix," *Statistics and Probability Letters*, vol. 73, no. 5, pp. 449–457, May 2005.

[3] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Trans. on Sig. Proc.*, vol. 56, no. 2, February 2008.

[4] N. Lu and D. Zimmerman, "On likelihood-based inference for a separable covariance matrix," Tech. Rep., Statistics and Actuarial Science Dept., Univ. of Iowa, Iowa City, IA, 2004.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] 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*, March 2008.

[7] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[8] Roger A. Horn and Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[9] Theodoros Tsiligkaridis and Alfred O. Hero III, "High dimensional covariance estimation under kronecker product structure," Tech. Rep., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 2011.

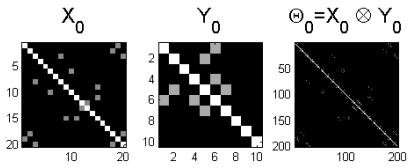

Fig. 1. Precision Matrices (Erdös-Rényi construction).

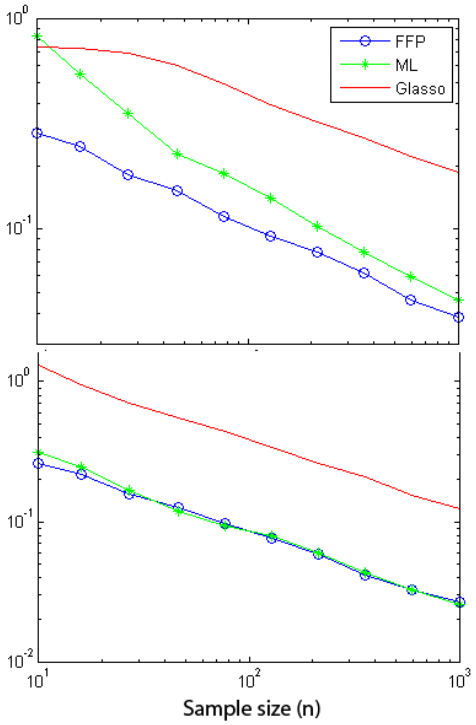

Fig. 2. Normalized rmse of precision matrix estimate (top) and covariance matrix estimate (bottom) as a function of sample size for structure exhibited in Fig. 1. Proposed FFP algorithm exhibits best performance.

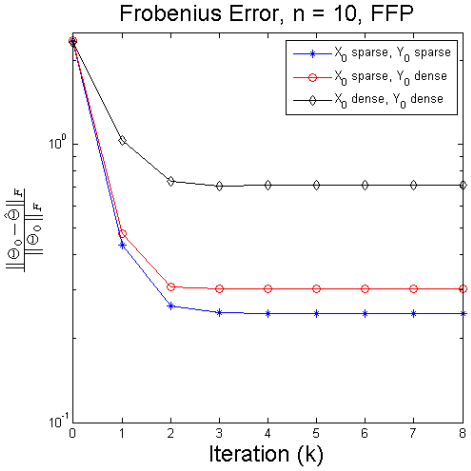

Fig. 3. Normalized rmse of FFP precision matrix estimate at successive iterations for case of both factors dense (top curve), only one factor sparse (middle curve), and both factors sparse (bottom curve). In all cases the FFP convergence rate is fast and the steady state rmse improves with sparsity.