High Dimensional Separable Representations for Statistical Estimation and Controlled Sensing

Theodoros Tsiligkaridis[†]

[†] Dept. of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor

Ph.D. Thesis Presentation
December 11, 2013

Motivation

- Separable approximations effective dimensionality reduction techniques for high dimensional problems.
- Covariance estimation: reduced computational complexity & improved estimation accuracy. Statistical estimation performance for separable models in high dimensions? Model mismatch?
- Centralized controlled sensing leads to great performance gains at the expense of query design. Separable approximations to optimal joint policy? Performance degradation?
- Controlled sensing over a network of greedy agents. Separable representation of information state? Separable representation of policy? Convergence?

Application: Spatiotemporal Signal Processing

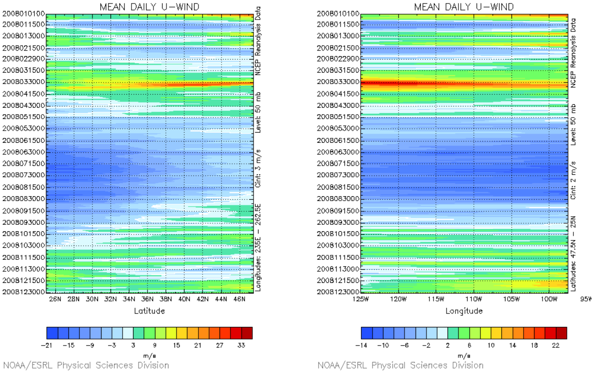


Figure : U-component of wind speed as a function of time and latitude/longitude for year 2008. (Source: National Centers for Environmental Prediction, NOAA)

Application: Centralized Active Multisensor Target Localization

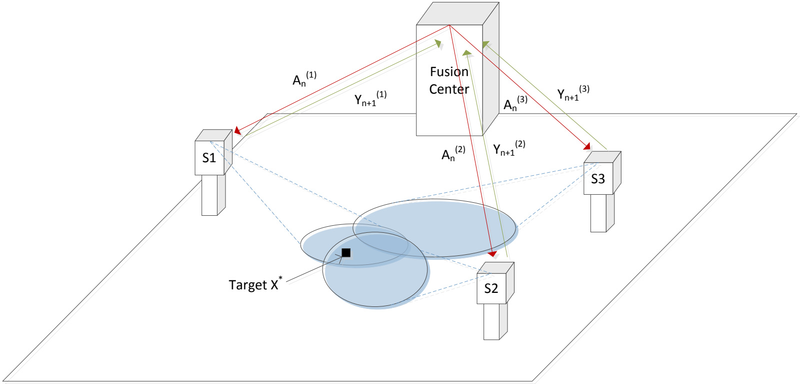


Figure : Illustration of basic centralized collaborative tracking system.

Application: Decentralized Active Multisensor Target Localization

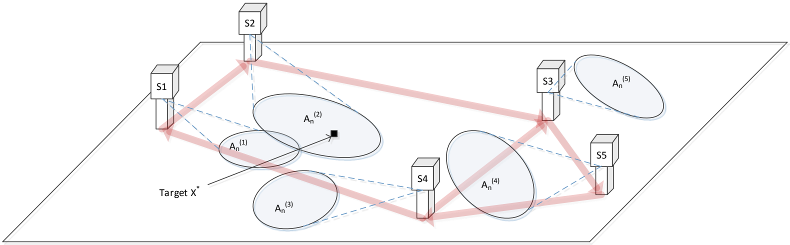


Figure : Illustration of basic decentralized collaborative tracking system.

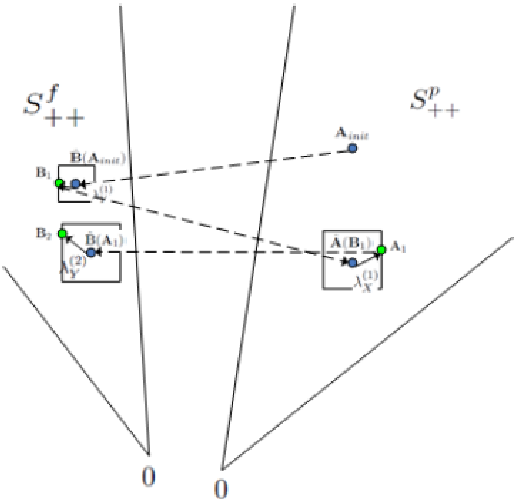
Impact

- **Engineering:** collaborative on-road vehicle-recognition & tracking, optimization & design of active sensing systems (e.g., frequency agile radar, multicamera object tracking with PTZ cameras), conditions on network structure for successful aggregation of information in decentralized settings, human-in-the-loop decision making
- **Signal Processing & Control:** covariance decompositions for multidimensional data with theoretical guarantees, centralized & decentralized collaborative estimation with active queries, non-Bayesian social learning with active queries over finite networks leads to global consistency, decentralized stochastic search
- **Social Sciences:** social learning & opinion dynamics, adaptive testing, recommendation systems, multitask learning, interview design

Contributions of Thesis

1. Performance bounds for high-dimensional Kronecker-product structured covariance matrix estimation
2. Optimal query design for a centralized collaborative controlled sensing system for target localization
3. Global convergence theory for decentralized collaborative controlled sensing for target localization

Kronecker Graphical Lasso



Mathematical setting

Observed $d \times n$ random matrix:

$$\mathbb{Z} = \begin{bmatrix} z_{1,1} & \cdots & z_{1,n} \\ \vdots & \ddots & \vdots \\ z_{d,1} & \cdots & z_{d,n} \end{bmatrix} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$$

Each column of $\mathbb{Z}$ is an independent realization of Gaussian random vector

$$\mathbf{z} = [z_1, \ldots, z_d]^T$$

Of interest: estimate the $d \times d$ inverse covariance (precision) matrix of $\mathbf{z}$ (and the covariance matrix)

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1}, \quad \mathbf{\Sigma} = \mathrm{cov}(\mathbf{z}) = E[\mathbf{z}\mathbf{z}^T]$$

Gaussian graphical models: activity recognition, gene expression networks, social networks, multiple financial time series.

Gaussian Graphical Models

Consider a random vector measurement $\mathbf{Z} \in \mathbb{R}^d$. Joint probability distribution of d measurements can be represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Edge $(i,j) \notin \mathcal{E}$ iff Z_i and Z_j are conditionally independent given all the other variables.

- If $\mathbf{Z}$ is a Gaussian random vector, conditional independence relationships between variables are encoded in precision matrix (Lauritzen [1996]). Thus, estimating the Gaussian graphical model is equivalent to estimating the precision matrix.
- Sparse GGM equivalent to sparse precision matrix.

Define sparsity parameter:

$$s_{\Theta_0} = \mathrm{card}\left(\{(i,j) : [\boldsymbol{\Theta}_0]_{i,j} \neq 0, i \neq j\}\right)$$

Sparse inverse covariance matrices and associated graphical models

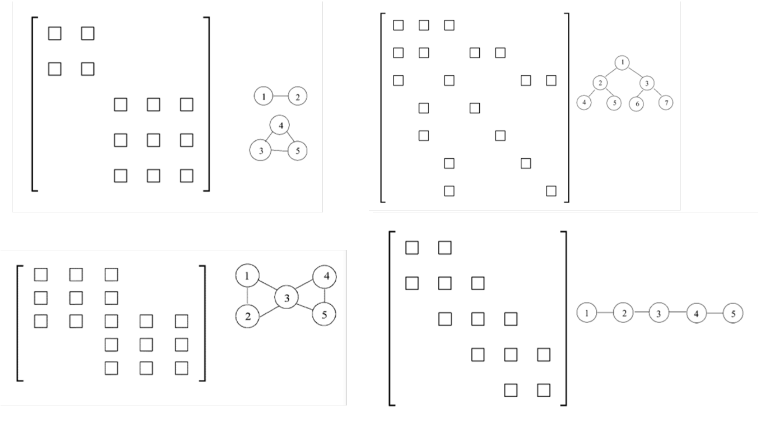


Figure : Left: inverse correlation matrix. Right: associated graphical model (Wiesel et al. [2010])

Prior Work

- Many more unknown parameters ($d(d+1)/2$) than measurements (n).
- Sample covariance matrix $\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t^T$ is poor estimator of $\mathbf{\Sigma}$:
  - Large eigenvalue spread in high dimensional regime (Karoui [2008]).
  - Estimation of eigenvectors of the SCM becomes impossible if the ratio n/d is below a critical threshold (Paul [2007], Rao et al. [2008]).
- Regularize:
  - Parametric models: Toeplitz, AR, ARMA (Bickel and Levina [2008], Huang et al. [2006], Cai et al. [2012]).
  - Sparse structured (inverse) covariance: Graphical lasso (Yuan and Lin [2007])
  - Kronecker structured covariance: Flip-Flop Kronecker covariance estimator (Werner et al. [2008])

Kronecker product model for covariance matrix

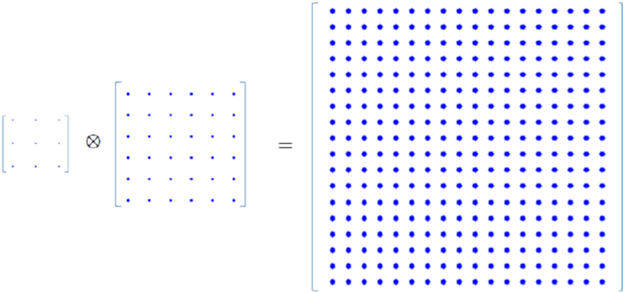


Figure : A saturated model with 18×18 covariance matrix has 18*(18+1)/2=171 unknown covariance parameters. A Kronecker product covariance model reduces number of parameters to $6 + 21 = 27$ unknown covariance parameters.

Sparse Kronecker product model for covariance matrix

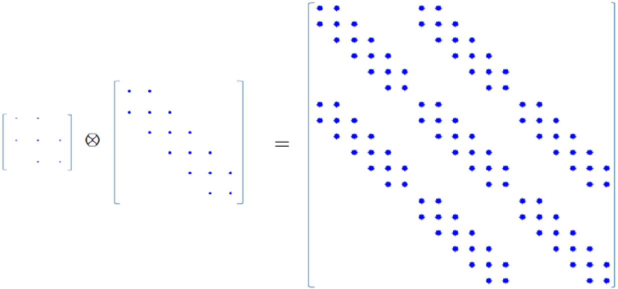


Figure : A sparse Kronecker product covariance model reduces number of parameters from 65 to 16 unknown covariance parameters.

Applications of KP Covariance

- geostatistics (Cressie [1993], Genton [2007])
- genomics (Yin and Li [2012])
- multi-task learning (Bonilla et al. [2008])
- face recognition (Zhang and Schneider [2010])
- recommendation systems (Allen and Tibshirani [2010])
- collaborative filtering (Yu et al. [2009])
- MIMO wireless communications (Werner and Jansson [2007])

Problem Formulation

- Available are n i.i.d. multivariate Gaussian observations $\{\mathbf{z}_t\}_{t=1}^n$, where $\mathbf{z}_t \in \mathbb{R}^{pq}$, having zero-mean and covariance equal to

$$\mathbf{\Sigma} = \underbrace{\mathbf{A}_0}_{p \times p} \otimes \underbrace{\mathbf{B}_0}_{q \times q} = \begin{bmatrix} [\mathbf{A}_0]_{1,1}\mathbf{B}_0 & \dots & [\mathbf{A}_0]_{1,p}\mathbf{B}_0 \\ \vdots & \ddots & \vdots \\ [\mathbf{A}_0]_{p,1}\mathbf{B}_0 & \dots & [\mathbf{A}_0]_{p,p}\mathbf{B}_0 \end{bmatrix},$$

  where $\mathbf{A}_0 \in S_{++}^p$ and $\mathbf{B}_0 \in S_{++}^q$.
- Goal is to estimate the covariance matrix and its inverse $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ (precision matrix).

Graphical Lasso (Yuan and Lin [2007])

Penalized negative log-likelihood function for $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$:

$$J(\mathbf{\Theta}) := \mathrm{tr}(\mathbf{\Theta}\hat{\mathbf{S}}_n) - \log\det(\mathbf{\Theta}) + \lambda|\mathbf{\Theta}|_1 \tag{1}$$

where $\hat{\mathbf{S}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{z}_t\mathbf{z}_t^T$ is the sample covariance matrix (SCM).
Minimizer $\hat{\mathbf{\Theta}}_n \in \arg\min J(\mathbf{\Theta})$.

- Fast algorithms exist for minimizing (1) (Friedman et al. [2008], Hsieh et al. [2011]) with worst-case computational complexity of $\mathcal{O}(d^4)$.
- High-dimensional MSE convergence rate (Rothman et al. [2008]):

$$\|\hat{\mathbf{\Theta}}_n - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(d + s_{\Theta_0})\log(d)}{n}\right) \tag{2}$$

where $\lambda \asymp \sqrt{\frac{\log(d)}{n}}$.

ML estimator of Kronecker structured covariance

Negative log-likelihood function when $\mathbf{\Theta}$ has Kronecker structure $\mathbf{\Theta} = \mathbf{X} \otimes \mathbf{Y}$:

$$J(\mathbf{X}, \mathbf{Y}) = \mathrm{tr}((\mathbf{X} \otimes \mathbf{Y})\hat{\mathbf{S}}_n) - q \log\det(\mathbf{X}) - p \log\det(\mathbf{Y}) \tag{3}$$

Alternating minimization yields Flip-Flop algorithm (Werner et al. [2008]) that generates updates of $\mathbf{A} = \mathbf{X}^{-1}$, $\mathbf{B} = \mathbf{Y}^{-1}$

$$\underbrace{\hat{\mathbf{A}}(\mathbf{B})}_{p \times p} = \frac{1}{q} \sum_{k,l=1}^{q} [\mathbf{B}^{-1}]_{k,l} \overline{\hat{\mathbf{S}}_n}(l, k) \tag{4}$$

$$\underbrace{\hat{\mathbf{B}}(\mathbf{A})}_{q \times q} = \frac{1}{p} \sum_{i,j=1}^{p} [\mathbf{A}^{-1}]_{i,j} \hat{\mathbf{S}}_n(j, i) \tag{5}$$

where $\overline{\hat{\mathbf{S}}}_n = \mathbf{K}_{p,q}^T \hat{\mathbf{S}}_n \mathbf{K}_{p,q}$
and $\mathbf{K}_{p,q} \mathrm{vec}(\mathbf{N}) = \mathrm{vec}(\mathbf{N}^T)$ for any $p \times q$ matrix $\mathbf{N}$.

Submatrix partitioning of SCM

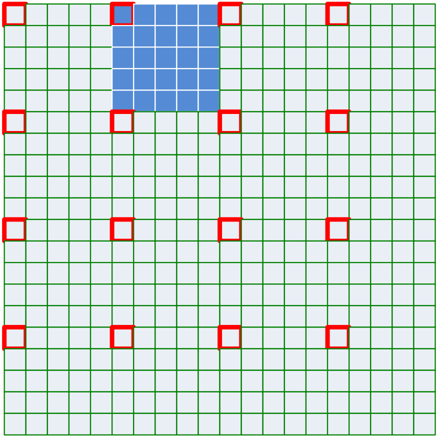


Figure : SCM of size $pq \times pq$ with $p = 4, q = 5$. Blue: $\hat{\mathbf{S}}_n(1,2)$. Red: $\overline{\hat{\mathbf{S}}}_n(1,1)$

MSE Convergence Rate of FF (Tsiligkaridis et al. [2012])

Let $\hat{\mathbf{R}}_{FF}(3) := \hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})) \otimes \hat{\mathbf{B}}(\hat{\mathbf{A}}(\hat{\mathbf{B}}(\mathbf{A}_{init})))$ denote the 3-step (noniterative) version of the flip-flop algorithm (Werner et al. [2008]). More generally, let $\hat{\mathbf{R}}_{FF}(k)$ denote the k-step version of the flip-flop algorithm.

Theorem

Let $\mathbf{A}_0, \mathbf{B}_0$, and $\mathbf{A}_{init}$ have uniformly bounded spectra and define $M = \max(p, f, n)$. Assume $p \geq q \geq 2$ and $p \log M \leq C'' n$ for some finite constant $C'' > 0$. Finally, assume $n \geq \frac{p}{q} + 1$. Then, for $k \geq 2$ finite,

$$\|\mathbf{\Theta}_{FF}(k) - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(p^2 + q^2)\log M}{n}\right) \tag{6}$$

as $n \to \infty$.

KGlasso Algorithm

$$\min J_\lambda(\mathbf{X}, \mathbf{Y}) = J(\mathbf{X}, \mathbf{Y}) + \lambda_X |\mathbf{X}|_1 + \lambda_Y |\mathbf{Y}|_1 \tag{7}$$

where $J(\cdot, \cdot)$ is given in (3) and $\lambda_X, \lambda_Y \geq 0$.

Algorithm 1 KGlasso (Tsiligkaridis et al. [2012, 2013a])

Input: $\hat{\mathbf{S}}_n$, p, q, n, $\lambda_X > 0$, $\lambda_Y > 0$
Output: $\hat{\mathbf{\Theta}}_{KGlasso}$
Initialize $\mathbf{A}_{init}$ to be positive definite.
$\hat{\mathbf{A}} \leftarrow \mathbf{A}_{init}$
repeat
    $\hat{\mathbf{B}} \leftarrow \frac{1}{p} \sum_{i,j=1}^{p} [\hat{\mathbf{A}}^{-1}]_{i,j} \hat{\mathbf{S}}_n(j, i)$
    $\check{\mathbf{Y}} \leftarrow \arg\min_{\mathbf{Y} \in S_{++}^q} \mathrm{tr}(\mathbf{Y}\hat{\mathbf{B}}) - \log\det(\mathbf{Y}) + \lambda_Y |\mathbf{Y}|_1$
    $\hat{\mathbf{A}} \leftarrow \frac{1}{q} \sum_{k,l=1}^{q} [\hat{\mathbf{B}}^{-1}]_{k,l} \overline{\hat{\mathbf{S}}}_n(l, k)$
    $\check{\mathbf{X}} \leftarrow \arg\min_{\mathbf{X} \in S_{++}^p} \mathrm{tr}(\mathbf{X}\hat{\mathbf{A}}) - \log\det(\mathbf{X}) + \lambda_X |\mathbf{X}|_1$
until convergence
$\hat{\mathbf{\Theta}}_{KGlasso} \leftarrow \check{\mathbf{X}} \otimes \check{\mathbf{Y}}$

Computational complexity: $\mathcal{O}(p^4 + q^4)$ (KGlasso)

KGlasso Convergence Rate (Tsiligkaridis et al. [2012])

Define $\mathbf{\Theta}_{KGlasso}(k)$ as the output of the kth KGlasso iteration.

Theorem

Let $\mathbf{A}_0, \mathbf{B}_0, \mathbf{A}_{init}$ have uniformly bounded spectra. Let $M = \max(p, f, n)$. Assume sparse $\mathbf{X}_0$ and $\mathbf{Y}_0$, i.e. $s_{X_0} = O(p), s_{Y_0} = O(f)$. Assume $\max\left(\frac{p}{q}, \frac{q}{p}\right) \log M = o(n)$. If in the KGlasso algorithm $\lambda_X^{(k)} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{q}}\right) q \sqrt{\frac{\log M}{n}}$ and $\lambda_Y^{(k')} \asymp \left(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{q}}\right) p \sqrt{\frac{\log M}{n}}$ for all $k, k' \geq 1$, then

$$\|\mathbf{\Theta}_{KGlasso}(k) - \mathbf{\Theta}_0\|_F^2 = O_P\left(\frac{(p+q)\log M}{n}\right) \tag{8}$$

as $n \to \infty$.

Assume $p \sim q$. Comparing the KGlasso convergence rate $(p+q)/n$ (8) with others

- SCM rate: $p^2 q^2/n$. Worse by 3 orders of magnitude
- FF rate: $(p^2 + q^2)/n$. Worse by 1 order of magnitude
- Glasso rate: $(pq + s_{\Theta_0})/n$. Worse by 1 order of magnitude.

Large Sample MSE Convergence

We considered $\mathbf{X}_0$ and $\mathbf{Y}_0$ large sparse matrices of dimension $p = q = 100$ yielding a covariance matrix $\mathbf{\Theta}_0$ of dimension $10,000 \times 10,000$. This dimension was too large for implementation of Glasso even when implemented using the state-of-the-art algorithm (Hsieh et al. [2011]). However, we can run KGlasso and FF and compare performances since they have considerably less computational burden.

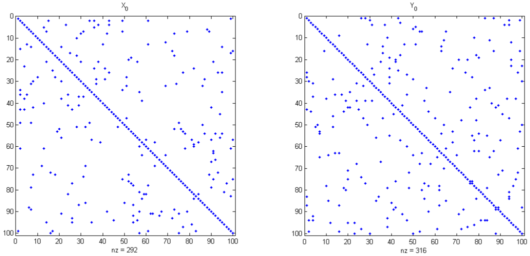


Figure : Sparse Kronecker matrix representation. Left panel: left Kronecker factor. Right panel: right Kronecker factor. The sparsity factor for both precision matrices is approximately 200.

Large Sample MSE Convergence (Cont.)

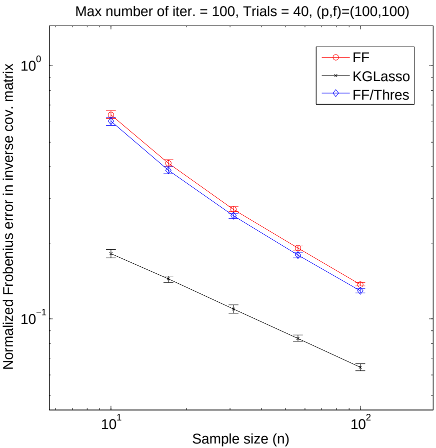
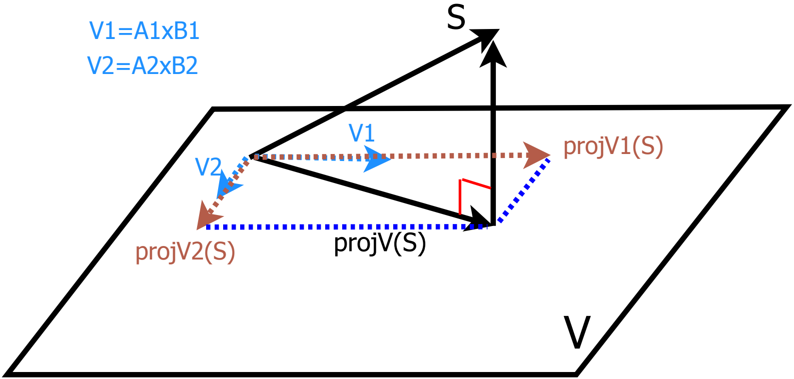


Figure : Normalized RMSE performance for precision matrix as a function of sample size n. For $n = 10$, there is a 72% RMSE reduction from the FF to KGLasso solution and a 70% RMSE reduction from the FF/Thres to KGLasso.

Kronecker PCA



Introduction

- Represent covariance as a Sum of Kronecker Products (SKP) of two lower dimensional factor matrices.

$$\mathbf{\Sigma}_0 = \sum_{\gamma=1}^{r} \mathbf{A}_{0,\gamma} \otimes \mathbf{B}_{0,\gamma} \tag{9}$$

where $\{\mathbf{A}_{0,\gamma}\}$ are $p \times p$ linearly independent matrices and $\{\mathbf{B}_{0,\gamma}\}$ are $q \times q$ linearly independent matrices.

- Note $1 \leq r \leq r_0 = \min(p^2, q^2)$ and refer to r as the *separation rank*.

Introduction

Applications of Sum of Kronecker Products (SKP) model (9)

- Spatiotemporal MEG/EEG covariance modeling (de Munck et al. [2002, 2004], Bijma et al. [2005], Jun et al. [2006])
- Synthetic Aperture Radar (SAR) data analysis (Tebaldini [2009], Rucci et al. [2010])

Van Loan and Pitsianis [1993]:

- Any $pq \times pq$ matrix $\mathbf{\Sigma}_0$ can be written as an orthogonal expansion of Kronecker products of the form (9)
- Low separation rank is *equivalent* to low rank in a permuted space defined by the reshaping operator $\mathcal{R}(\cdot)$

Low separation rank $\Leftrightarrow$ Low rank in permuted space

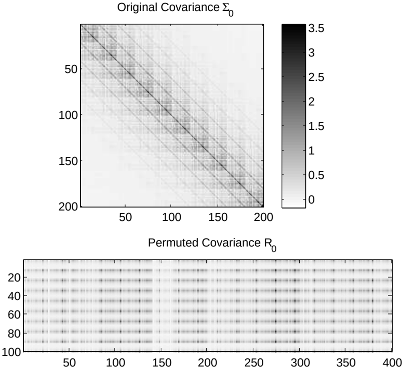


Figure : Original (top) and permuted covariance (bottom) matrix. The original covariance is $\mathbf{\Sigma}_0 = \mathbf{A}_0 \otimes \mathbf{B}_0$, where $\mathbf{A}_0$ is a 10×10 Toeplitz matrix and $\mathbf{B}_0$ is a 20×20 unstructured p.d. matrix. Note that the permutation operator $\mathcal{R}$ maps a symmetric p.s.d. matrix $\mathbf{\Sigma}_0$ to a non-symmetric rank 1 matrix $\mathbf{R}_0 = \mathcal{R}(\mathbf{\Sigma}_0)$.

Permuted rank-penalized least-squares (PRLS) (Tsiligkaridis and Hero [2013a,b])

1. Map SCM to a different linear space:

$$\hat{\mathbf{R}}_n = \mathcal{R}(\hat{\mathbf{S}}_n) \in \mathbb{R}^{p^2 \times q^2}$$

2. Solve least-squares problem with nuclear norm penalization:

$$\hat{\mathbf{R}}_n^\lambda \in \arg \min_{\mathbf{R} \in \mathbb{R}^{p^2 \times q^2}} \|\hat{\mathbf{R}}_n - \mathbf{R}\|_F^2 + \lambda \|\mathbf{R}\|_* \tag{10}$$

3. Map back to original space:

$$\hat{\mathbf{S}}_n^\lambda = \mathcal{R}^{-1}(\hat{\mathbf{R}}_n^\lambda) \in \mathbb{R}^{pq \times pq}$$

where $\lambda \geq 0$ is a regularization parameter.

Properties of PRLS Estimator (Tsiligkaridis and Hero [2013b])

Theorem

- *The solution* $\hat{\mathbf{\Sigma}}_n^\lambda$ *is symmetric.*
- *If* $n \geq pq$, *then the solution* $\hat{\mathbf{\Sigma}}_n^\lambda$ *is positive definite with probability 1.*

Theorem

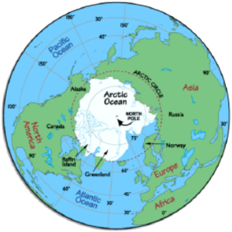
Define $M = \max(p, q, n)$. *Set*
$\lambda = \lambda_n = \frac{2C_0 t}{1-2\epsilon'} \max\left\{\frac{p^2+q^2+\log M}{n}, \sqrt{\frac{p^2+q^2+\log M}{n}}\right\}$ *for* $t > 0$ *large enough.*
Then, with probability at least $1 - 2M^{-\frac{t}{4C}}$:

$$\|\hat{\mathbf{\Sigma}}_n^\lambda - \mathbf{\Sigma}_0\|_F^2 \leq \inf_{\mathbf{R}:rank(\mathbf{R})\leq r} \|\mathbf{R} - \mathbf{R}_0\|_F^2 + C' r \max\left\{\left(\frac{p^2+q^2+\log M}{n}\right)^2, \frac{p^2+q^2+\log M}{n}\right\} \quad (11)$$

for some absolute constant $C' > 0$.

Setup

- NCEP Dataset: Daily average wind speeds collected at $q = 144 \times 73$ weather stations spread throughout the world (Kalnay et al. [1996], Tsiligkaridis and Hero [2013b])
- Considered a 10×10 grid of stations, corresponding to latitude range 90°N-67.5°N and longitude range 0°E-22.5°E



- Prediction time lag $p - 1 = 7$, full dimension $d = pq = 800$, number of training samples $n = 228$.
- Training period: $2003 - 2007$, Testing period: $2008 - 2012$.

Kronecker product decomposition: PRLS

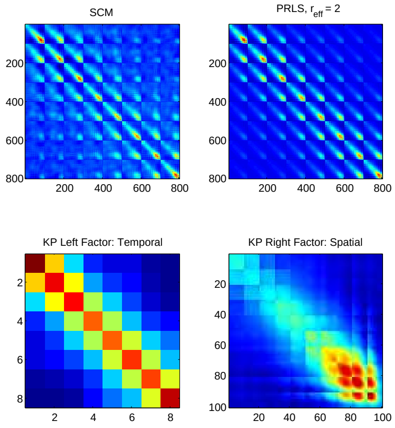


Figure : Sample covariance matrix (SCM) (top left), PRLS covariance estimate (top right), temporal Kronecker factor for first KP component (bottom left) and spatial Kronecker factor for first KP component (bottom right).

Kronecker Spectrum

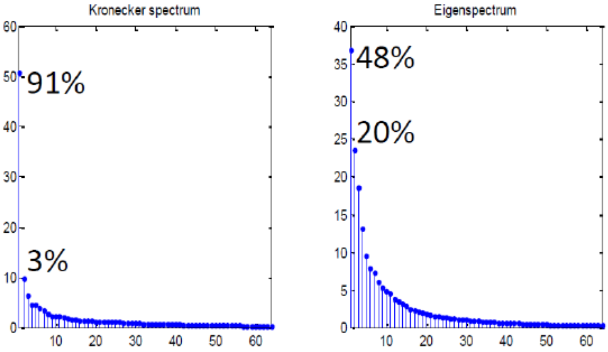


Figure : Kronecker spectrum of SCM (left) and Eigenspectrum of SCM (right). The KP spectrum is more compact than the eigenspectrum.

RMSE performance gains

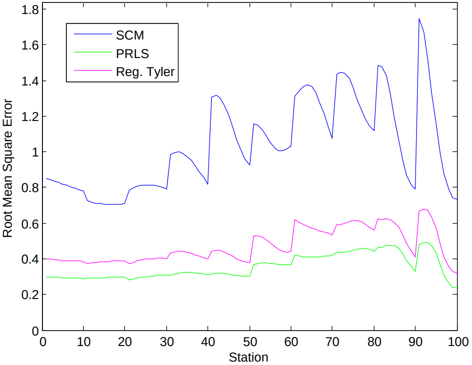
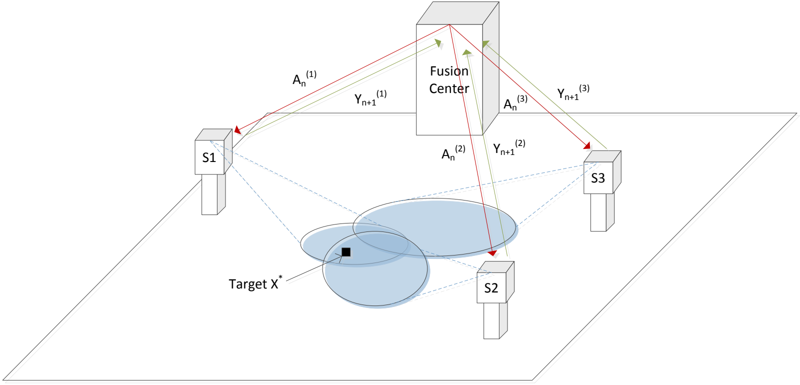


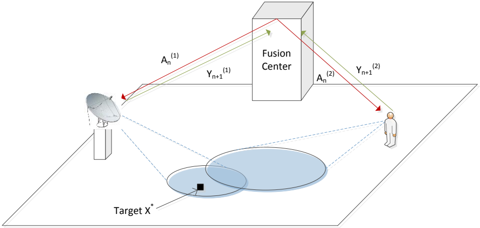
Figure : RMSE prediction performance across q stations for linear estimators using SCM (blue), PRLS (green) and regularized Tyler (magenta).

- Average gain of PRLS over SCM $= 4.64$ dB
- Average gain of Reg. Tyler over SCM $= 3.41$ dB

Centralized Collaborative 20 Questions



Motivation



- What is the intrinsic value of adding a human-in-the-loop to an autonomous learning machine?
- Insight into human-aided autonomous sensing for estimating an unknown target location or identifying a target.

Motivation

Figure : PTZ IP camera. Source: en.wikipedia.org/wiki/Pan-tilt-zoom_ camera

- Sensor systems become more flexible, e.g. pan-tilt-zoom cameras: where to look? different sensor waveforms & observations modes? How to control these aspects for a common localization objective?

Prior Work & Applications

Ask a sequence of questions and refine posterior distribution of target's location given the responses.

- Probabilistic Bisection Algorithm (PBA) first introduced in (Horstein [1963]).
- Discretized PBA (Burnashev and Zigangirov [1974]).
- Noisy Binary Search (Karp and Kleinberg [2007]).
- Convergence rate for BZ algorithm (Castro and Nowak [2007]).
- Noisy 20 questions game: PBA shown to be optimal under minimum expected entropy criterion (Jedynak et al. [2012]).
- Convergence rate for PBA (Waeber et al. [2013]).

Applications of PBA: stochastic root finding, combinatorial optimization, road tracking, electron microscopy

Single player setting

- Jedynak et al. [2012] considers 20 questions with noise, where a noisy oracle is queried whether a target X^* lies in a set $A_n \subset \mathbb{R}^d$.
- Starting with a prior distribution on the target's location $p_0(\cdot)$, minimize expected entropy of the posterior distribution:

$$\inf_{\pi} \mathrm{E}^{\pi} \left[H(p_N) \right] \tag{12}$$

  where $\pi = (\pi_0, \pi_1, \ldots)$ denotes the policy. The posterior mean/median of $p_N(\cdot)$ is the target location estimate.
- Jedynak et al. [2012] shows the bisection policy is optimal under the minimum entropy criterion. Assuming the noisy channel is a BSC, optimal policies are characterized by:

$$\mathbb{P}_n(A_n) := \int_{A_n} p_n(x)dx = 1/2 \tag{13}$$

Noisy 20 Questions with Collaborative Players: Model (Tsiligkaridis et al. [2013c])

- M collaborating players can be asked questions at each time instant.
- mth player's query at time n: "does X^* lie in the region $A_n^{(m)} \subset \mathbb{R}^d$?"
- Query is the binary variable $Z_n^{(m)} = I(X^* \in A_n^{(m)}) \in \{0, 1\}$ to which the player yields provides a noisy response $Y_{n+1}^{(m)} \in \{0, 1\}$.
- Define the M-tuples $\mathbf{Y}_{n+1} = (Y_{n+1}^{(1)}, \ldots, Y_{n+1}^{(M)})$ and $\mathbf{A}_n = \{A_n^{(1)}, \ldots, A_n^{(M)}\}$.

Assumption

Players' responses are conditionally independent:

$$\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y} | \mathbf{A}_n, X^* = x, \mathcal{F}_n) = \prod_{m=1}^{M} \mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \mathcal{F}_n) \quad (14)$$

$$\mathbb{P}(Y_{n+1}^{(m)} = y^{(m)} | A_n^{(m)}, X^* = x, \mathcal{F}_n) = \begin{cases} f_1^{(m)}(y^{(m)} | \epsilon_m), & x \in A_n^{(m)} \\ f_0^{(m)}(y^{(m)} | \epsilon_m), & x \notin A_n^{(m)} \end{cases} \quad (15)$$

$$f_j^{(m)}(y^{(m)} | \epsilon_m) = \begin{cases} 1 - \epsilon_m, & y^{(m)} = j \\ \epsilon_m, & y^{(m)} = 1 - j \end{cases} \quad (16)$$

Optimal Joint Query Design: Setup

- Joint controller chooses M queries $A_n^{(m)}$ at time n. Define the set of subsets of $\mathbb{R}^d$:

$$\gamma(A^{(1)}, \dots, A^{(M)}) = \left\{ \bigcap_{m=1}^{M} (A^{(m)})^{i_m} : i_m \in \{0,1\} \right\}$$

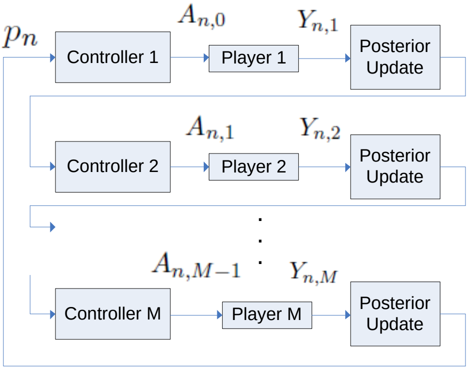
  where $(A)^0 := A^c$ and $(A)^1 := A$. The cardinality of this set of subsets is 2^M and these subsets partition $\mathbb{R}^d$.

- Define the density parameterized by $\mathbf{A}_n, p_n, i_1, \dots, i_M$:

$$g_{i_1:i_M}(y^{(1)}, \dots, y^{(M)} | \mathbf{A}_n, \mathcal{F}_n) := \prod_{m=1}^{M} f_{i_m}^{(m)}(y^{(m)} | A_n^{(m)}, \mathcal{F}_n)$$

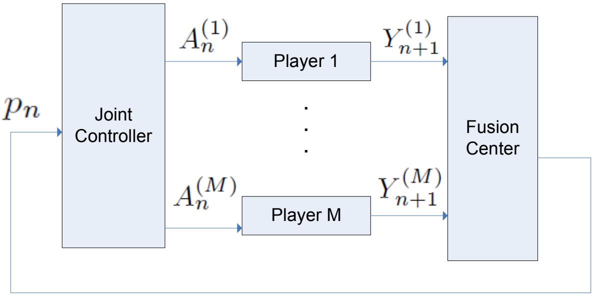
  where $i_j \in \{0,1\}$.

Sequential Query Design



- Query region A_{n_t} chosen at time $n_t = (n, t)$, where $n = 0, 1, \ldots$ indexes over cycles and $t = 0, \ldots, M-1$ indexes within cycles.
- Nested sequence of sigma-algebras $\mathcal{G}_{n,t}$, $\mathcal{G}_{n,t} \subset \mathcal{G}_{n+i,t+j}$ for all $i \geq 0$ and $j \in \{0, \ldots, M-1-t\}$, generated by sequence of queries and the players' responses.

Optimal Joint Query Design



- Joint controller chooses a batch of M queries $\{A_n^{(m)}\}$ at time n.
- As in sequential query design, joint queries chosen based on accumulation information at controller. Since full batch of joint queries are determined at start of nth cycle, the joint controller only has access to a coarser filtration $\mathcal{F}_n$, $\mathcal{F}_{n-1} \subset \mathcal{F}_n$, as compared to $\mathcal{G}_{n,t}$.

Equivalence Theorem (Tsiligkaridis et al. [2013c])

Theorem

(Equivalence, Known Error Probabilities)

1. *The expected entropy loss under an optimal joint query design is the same as the greedy sequential query design. This loss is given by:*

$$C = \sum_{m=1}^{M} C(\epsilon_m) = \sum_{m=1}^{M} (1 - h_b(\epsilon_m)) \tag{17}$$

where $h_b(\epsilon_m) = -\epsilon_m \log(\epsilon_m) - (1-\epsilon_m)\log(1-\epsilon_m)$ *is the binary entropy function.*

2. *All jointly optimal control laws equalize the posterior probability over the dyadic partitions induced by* $\mathbf{A}_n = \{A_n^{(1)}, \ldots, A_n^{(M)}\}$:

$$\mathbb{P}_n(R) = \int_R p_n(x)dx = 2^{-M}, \forall R \in \gamma(\mathbf{A}_n). \tag{18}$$

Consequences of Equivalence Theorem

- Optimal policy can be implemented using the simpler sequential query design.
- Despite the fact that all players are conditionally independent, the joint policy does not decouple into separate single player optimal policies (analogous to the non-separability of the optimal vector-quantizer in source coding even for independent sources Gersho and Gray [1992]).
- Optimal queries must be overlapping-i.e., $\bigcap_{m=1}^{M} A_n^{(m)} \neq \emptyset$, but not identical.
- Optimal query $\mathbf{A}_n$ is not unique.

Example of optimal queries for $M = 2$

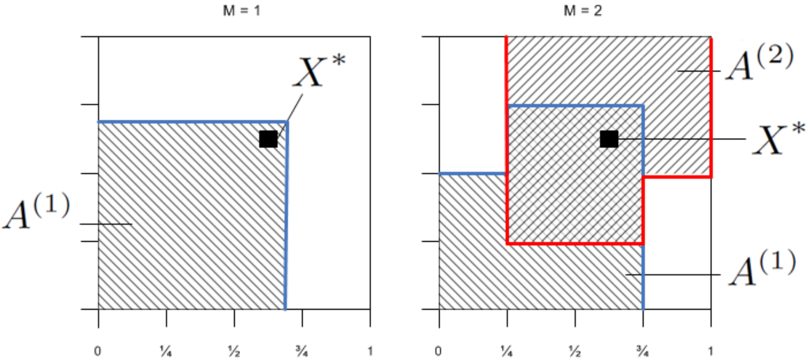


Figure : Jointly optimal queries under uniform prior.

Lower Bounds on MSE via Entropy Loss

Theorem

(Lower Bound on MSE) Assume the entropy $H(p_0)$ is finite. Then, the MSE of the joint or sequential query policies satisfies:

$$\frac{K}{2\pi e} d \exp\left(-\frac{2nC}{d}\right) \leq \mathbb{E}[\| X^* - X_n \|_2^2] \tag{19}$$

where $K = e^{2H(p_0)}$ and X_n is the posterior mean. The expected entropy loss per iteration is $C = \sum_m C(\epsilon_m)$.

Upper Bounds on MSE: Setup

- Performance analysis of PBA is difficult primarily due to the continuous nature of the posterior Castro and Nowak [2007].

  "The probabilistic bisection algorithm seems to work extremely well in practice, but it is hard to analyze and there are few theoretical guarantees for it, especially pertaining error rates of convergence."

- A discretized version of PBA was proposed in (Burnashev and Zigangirov [1974]) (BZ algorithm), which imposes a piecewise constant structure on the posterior (see Castro and Nowak [2007], App. A in Castro [2007]).
- Recently, an answer for the continuous PBA was given in (Waeber et al. [2013]) for one-dimensional target search.

Upper Bounds on MSE: Setup

- For simplicity, assume the target location is constrained to the unit interval $\mathcal{X} = [0, 1]$.
- A step size $\Delta > 0$ is defined such that $\Delta^{-1} \in \mathbb{N}$ and the posterior after j iterations is $p_j : \mathcal{X} \to \mathbb{R}$, given by

$$p_j(x) = \frac{1}{\Delta} \sum_{i=1}^{\Delta^{-1}} a_i(j) I(x \in I_i)$$

where $I_1 = [0, \Delta], I_i = ((i-1)\Delta, i\Delta]$ for $i = 2, \ldots, \Delta^{-1}$. The initial posterior is $a_i(0) = \Delta$. The posterior is characterized completely by the pseudo-posterior $\mathbf{a}(j) = [a_1(j), \ldots, a_{\Delta^{-1}}(j)]$ which is updated at each iteration via Bayes rule.

Upper Bounds on MSE

Theorem

(Upper Bound on MSE) Consider the sequential bisection algorithm for M players in one-dimension, where each bisection is implemented using the BZ algorithm. Then, we have:

$$\mathbb{P}(|X^* - \hat{X}_n| > \Delta) \leq (\frac{1}{\Delta} - 1) \exp\left(-n\bar{C}\right)$$

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq (2^{-2/3} + 2^{1/3}) \exp\left(-\frac{2}{3}n\bar{C}\right) \tag{20}$$

where $\bar{C} = \sum_{m=1}^{M} \bar{C}(\epsilon_m)$, $\bar{C}(\epsilon) = 1/2 - \sqrt{\epsilon(1-\epsilon)}$.

Upper Bounds on MSE: Human-in-the-loop

- Player 1 (machine) has constant error probability $\epsilon_1 \in (0, 1/2)$
- Player 2 (human) has error probability depending on the target localization error:

$$\mathbb{P}(Y_{n+1}^{(2)} = y^{(2)} | Z_n^{(2)} = 1 - y^{(2)}) = \frac{1}{2} - \min(\delta_0, \mu |X^* - X_n|^{\kappa - 1}) \tag{21}$$

- κ = human "resolution" ($\kappa > 1$)
- δ_0 = reliability parameter ($0 < \delta_0 < \mu < 1/2$)

- MSE upper bound for "player 1 + human" system:

$$\mathbb{E}[(X^* - \hat{X}_n)^2] \leq e^{-\frac{2}{3} n \bar{C}(\epsilon_1)} \times \left[2^{-2/3} + 2^{1/3} \exp\left(-\frac{\mu^2}{50} \left(\frac{3 \cdot 2^{-1/3}}{4} \right)^{2\kappa - 2} n e^{-n \bar{C}(\epsilon_1) \frac{2\kappa - 2}{3}} \right) \right] \tag{22}$$

which is no greater than the "player 1" MSE bound.

- Both bounds converge to zero at the same rate as $n \to \infty$.
- Human gain ratio (HGR) = ratio of MSE upper bounds associated with "player 1" and "player 1 + human".

$$R_n(\kappa) = \frac{2^{-2/3} + 2^{1/3}}{2^{-2/3} + 2^{1/3} \exp\left(-\frac{\mu^2}{50} \left(\frac{3 \cdot 2^{-1/3}}{4} \right)^{2\kappa - 2} n e^{-n \bar{C}(\epsilon_1) \frac{2\kappa - 2}{3}} \right)} \tag{23}$$

Upper Bounds on MSE: Human-Gain Ratio

- The larger ϵ_1 is, the larger is the HGR.
- As κ decreases to 1, the ratio increases, meaning that the human becomes more like the machine and helps more.

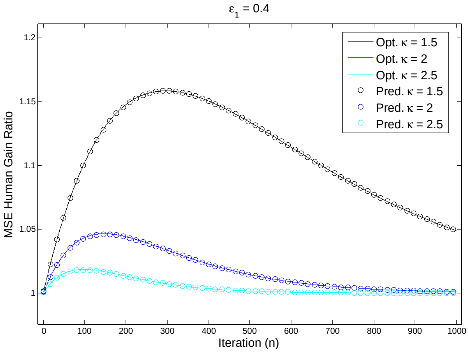


Figure : Human gain ratio as a function of κ. The human provides the largest gain in the beginning few iterations and its value of information decreases as $n \to \infty$. The predictions well match the optimized bounds.

Simulation (Known error probabilities): Initial Distribution

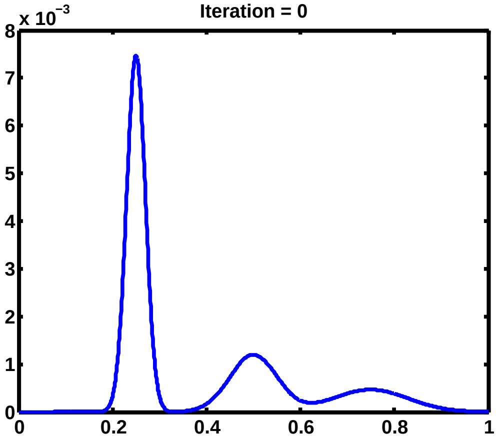


Figure : Initial distribution is a mixture of three Gaussians with means 0.25, 0.5 and 0.75, and variances 0.02, 0.05 and 0.08, respectively. The target was set to be the center of the mode at $X^* = 0.75$ with the largest variance.

Simulation (Known error probabilities): MSE Decay

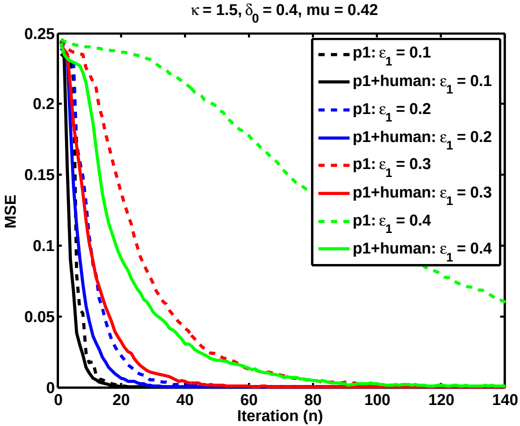
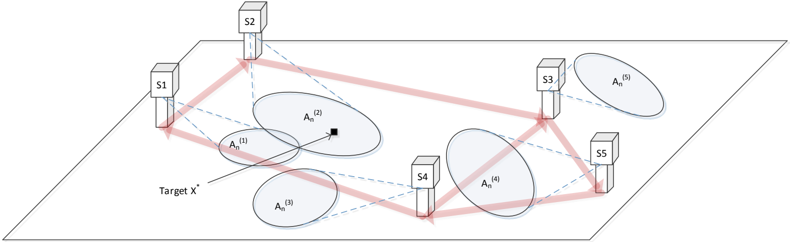


Figure : Monte Carlo simulation for MSE performance of the sequential estimator as a function of iteration and $\epsilon_1 \in (0, 1/2)$. 2000 Monte Carlo trials were used. The human parameters were set to $\kappa = 1.5, \mu = 0.42, \delta_0 = 0.4$, the length of pseudo-posterior was $\Delta^{-1} = 1618$. The target was set to $X^* = 0.75$.

Decentralized Collaborative 20 Questions



Motivation

Consider a collection of agents in a network with the objective of localizing a target collectively.

- What is the value of collaboration when there is no central authority?
- Local in-network querying and processing leads to global equilibrium? Deterministic or random limit? Unbiasedness?

Intractability of fully Bayesian methodology

- limited observability (observations of an agent not observable by others) & lack of global knowledge of observation statistics
- if agents have only partial information on the network structure and the probability distribution of the signals observed by other agents, the Bayesian approach becomes more complicated because agents would need to form and update beliefs on the states of the world, in addition to the networks struture and the rest of the agents' signal structures
- even if the network structure is known, agents would still need to update beliefs on the information of every other agent in the network, given only the neighbors' beliefs at each iteration

Prior Work on Distributed Averaging

Consensus, gossip algorithms, distributed averaging: messages distributed around network through local processing.

- averaging under randomized gossip (Boyd et al. [2006])
- geographic gossip (Dimakis et al. [2006])
- randomized path averaging (Benezit et al. [2010])
- gossip algorithms for sensor networks (Dimakis et al. [2010])
- randomized gossip broadcast algorithms for consensus (Aysal et al. [2009])
- gossip distributed estimation for linear parameter estimation (Kar and Moura [2011])
- consensus for wireless medium (Nokleby et al. [2013])

Applications: distributed optimization (Tsitsiklis [1984], Tsitsiklis et al. [1986]), load-balancing (Cybenko [1989]), distributed detection (Saligrama et al. [2006])

Our work differs because we consider new information injected into the dynamical system described by averaging and because we consider controlled observations.

Prior Work on Social Learning

Dynamic model of opinion formation.

- opinion formation model (DeGrout [1974])
- convergence of dynamics generated by non-Bayesian decentralized estimation scheme (Jadbabaie et al. [2012])
- rate of convergence analysis (Molavi et al. [2013])

Our work differs because we consider continuous-valued target space and controlled observations.

Prior Work on Computerized Adaptive Testing

Given current estimate of proficiency, how to choose next test item?

- dynamic selection of test items via item-response theory & maximum information or maximum expected precision criterion (Wainer [2000], Owen [1975])

Our work differs because we consider continuous-valued query regions, no practical constraints necessary, and a different objective function.

Prior Work on Active Stochastic Search/20 Questions

Active querying for sequential estimation.

- single-player 20 questions for target localization (Jedynak et al. [2012])
- convergence rate for discretized version of single-player 20 questions (Castro and Nowak [2007])
- convergence rate for continuous-space single-player PBA (Waeber et al. [2013])
- (centralized) multi-player 20 questions for target localization (Tsiligkaridis et al. [2013b])

Our work differs because we consider intermediate local belief sharing between agents after each local bisection and Bayesian update (entropy no longer monotonically decreasing for each agent!). Also, each agent incorporates beliefs of neighbors in a way that is agnostic of neighbors' error probabilities.

Notation

- $X^* \in [0,1]$ = true target location
- $\mathcal{N} = \{1, \ldots, M\}$ = agent set of network
- $G = (\mathcal{N}, E)$ directed graph capturing agent interactions
- $\mathcal{N}_i = \{j \in \mathcal{N} : (j,i) \in E\}$ = local neighborhood of ith agent
- $p_{i,t}(\cdot)$ = belief of ith agent at time t

Decentralized Estimation

Algorithm 2 Decentralized Estimation Algorithm

Input: $G = (\mathcal{N}, E), A = \{a_{i,j} : (i,j) \in \mathcal{N} \times \mathcal{N}\}, \{\epsilon_i : i \in \mathcal{N}\}$
Output: $\{\hat{X}_{i,t}, \check{X}_{i,t} : i \in \mathcal{N}\}$
Initialize $p_{i,0}(\cdot)$ to be positive everywhere.
repeat
  For each agent $i \in \mathcal{N}$:
    Bisect posterior density at median: $\hat{X}_{i,t} = F^{-1}_{i,t}(1/2)$.
    Obtain (noisy) binary response $y_{i,t+1} \in \{0,1\}$.
    Belief update:

$$p_{i,t+1}(x) = a_{i,i} p_{i,t}(x) \frac{l_i(y_{i,t+1}|x, \hat{X}_{i,t})}{\mathcal{Z}_{i,t}(y_{i,t+1})} + \sum_{j \in \mathcal{N}_i} a_{i,j} p_{j,t}(x), \qquad x \in \mathcal{X} \tag{24}$$

  where the observation p.m.f. is:

$$l_i(y|x, \hat{X}_{i,t}) = f_1^{(i)}(y) I(x \le \hat{X}_{i,t}) + f_0^{(i)}(y) I(x > \hat{X}_{i,t}), \qquad y \in \mathcal{Y} \tag{25}$$

  and $f_1^{(i)}(y) = (1-\epsilon_i)^{I(y=1)} \epsilon_i^{I(y=0)}$, $f_0^{(i)}(y) = 1 - f_1^{(i)}(y)$.
    Calculate target estimate: $\check{X}_{i,t} = \int_{\mathcal{X}} x p_{i,t}(x) dx$.
until convergence

Assumptions

- (Conditional Independence) Assume conditional independence:

$$\mathbb{P}(\mathbf{Y}_{t+1} = \mathbf{y}|\mathcal{F}_t) = \prod_{i=1}^{M} \mathbb{P}(Y_{i,t+1} = y_i|\mathcal{F}_t) \tag{26}$$

  and each player's response is governed by:

$$l_i(y_i|x, A_{i,t}) := \mathbb{P}(Y_{i,t+1} = y_i|A_{i,t}, X^* = x) = \begin{cases} f_1^{(i)}(y_i), & x \in A_{i,t} \\ f_0^{(i)}(y_i), & x \notin A_{i,t} \end{cases} \tag{27}$$

- (Memoryless Binary Symmetric Channels) Model players' responses as independent BSC's with crossover probabilities $\epsilon_i \in (0, 1/2)$.

$$f_z^{(i)}(y_i) = \begin{cases} 1 - \epsilon_i, & y_i = z \\ \epsilon_i, & y_i \neq z \end{cases}$$

  for $i = 1, \ldots, M, z = 0, 1$.

- (Strong Connectivity & Positive Self-reliances) Assume that the network is strongly connected and all self-reliances $a_{i,i}$ are strictly positive.

Global Convergence Theory

Theorem

(Asymptotic Agreement/Consensus) Consider Algorithm 2. Let $B = [0, b] \in \mathcal{B}([0, 1])$. Then, consensus of the agents' beliefs is asymptotically achieved across the network:

$$V_t(B) = \max_i \mathbb{P}_{i,t}(B) - \min_i \mathbb{P}_{i,t}(B) \xrightarrow{p.} 0$$

as $t \to \infty$.

Theorem

(Convergence of Beliefs to a Deterministic Limit & Consistency) Consider Algorithm 2. Let $B = [0, b] \in \mathcal{B}([0, 1])$. Then, we have:

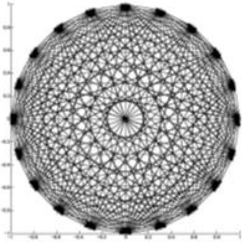
1. *For each $i \in \mathcal{N}$:*

$$F_{i,t}(b) = \mathbb{P}_{i,t}(B) \xrightarrow{p.} F_\infty(b) = \begin{cases} 0, & b < X^* \\ 1, & b > X^* \end{cases}$$

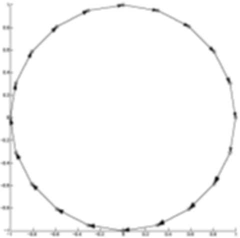
2. *For all $i \in \mathcal{N}$:*

$$\check{X}_{i,t} := \int_{x=0}^{1} x p_{i,t}(x) dx \xrightarrow{p.} X^*$$

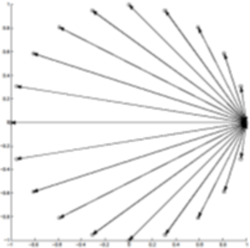
Simulation: Three network topologies



a) Fully connected graph

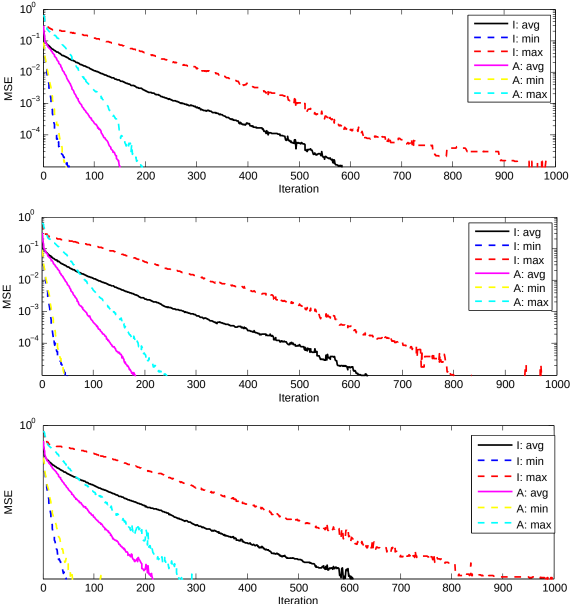


b) Cyclic graph

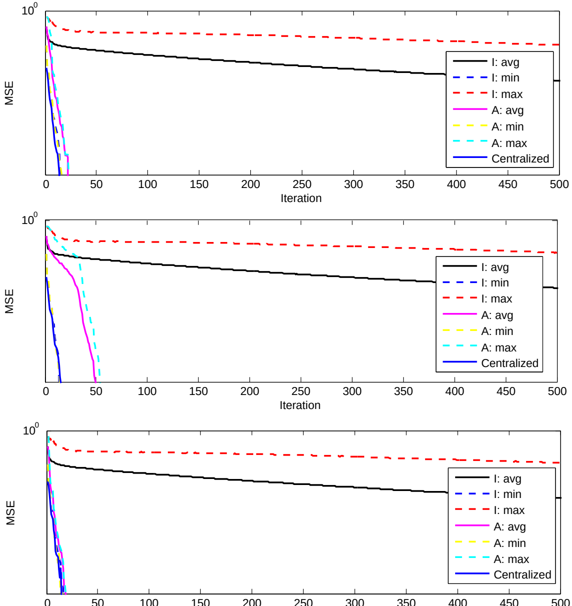


c) Star graph

MSE Performance, $\epsilon_i = 0.4, \forall i$



MSE Performance, $\epsilon_1 = 0.05, \epsilon_i = 0.45, \forall i \neq 1$



Main Contributions

1. Kronecker Graphical Lasso
   - Sparse covariance estimation algorithm (KGlasso) introduced for the high-dimensional setting for Kronecker product structure.
   - High-dimensional MSE convergence rate analysis.
   - Analysis prescribes selection of regularization parameters.
2. Covariance Estimation via Kronecker Product Expansions
   - Scalable covariance estimation algorithm (PRLS) introduced for the high-dimensional setting.
   - Tradeoff between approximation error and estimation error.
   - High-dimensional MSE convergence rate analysis.
   - Analysis prescribes selection of regularization parameter.

Main Contributions

3. Centralized Collaborative 20 Questions
    - Introduced model for centralized collaborative 20 questions.
    - Characterized optimal policies & proved equivalence theorem that simplifies policy implementation.
    - Incorporated human-in-the-loop by treating him as a collaborative player.
    - Linked information theoretic gains to MSE convergence rates.
4. Decentralized Collaborative 20 Questions
    - Introduced model for decentralized collaborative 20 questions.
    - Proved consensus of agents' beliefs & global consistency of decentralized estimation algorithm.

Thank you!

G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, 2010.

T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Transactions on Signal Processing*, 57(7), July 2009.

F. Benezit, A. Dimakis, P. Thiran, and M. Vetterli. Order-optimal consensus through randomized path averaging. *IEEE Transactions on Information Theory*, 56(10):5150–5167, October 2010.

P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604, 2008.

Fetsje Bijma, Jan de Munck, and Rob Heethaar. The spatiotemporal meg covariance matrix modeled as a sum of kronecker products. *NeuroImage*, 27:402–415, 2005.

E. Bonilla, K. M. Chai, and C. Williams. Multi-task gaussian process prediction. *Advances in Neural Information Processing Systems*, pages 153–160, 2008.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized Gossip Algorithms. *IEEE Transactions on Information Theory*, 52(6): 2508–2530, June 2006.

M. V. Burnashev and K. Sh. Zigangirov. An interval estimation problem for controlled observations. *Problems in Information Transmission*, 10: 223–231, 1974.

T. Tony Cai, Z. Ren, and H. Zhou. Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields*, March 2012.

R. Castro. *Active Learning and Adaptive Sampling for Non-parametric Inference*. PhD thesis, Rice University, August 2007.

R. Castro and R. Nowak. Active learning and sampling. In *Foundations and Applications of Sensor Management*. Springer, 2007.

N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

G. Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of Parallel and Distributed Computing*, 7(2): 279–301, 1989.

J. C. de Munck, H. M. Huizenga, L. J. Waldorp, and R. M. Heethaar. Estimating stationary dipoles from meg/eeg data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7), July 2002.

J. C. de Munck, F. Bijma, P. Gaura, C. A. Sieluzycki, M. I. Branco, and R. M. Heethaar. A maximum-likelihood estimator for trial-to-trial variations in noisy meg/eeg data sets. *IEEE Transactions on Biomedical Engineering*, 51(12), 2004.

M. H. DeGrout. Reaching a consensus. *Journal of American Statistical Association*, 69:118–121, 1974.

A. Dimakis, A. Sarwate, and M. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3):1205–1216, March 2006.

A. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11), November 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

M. G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18:681–695, 2007.

A. Gersho and R. M. Gray. *Vector quantization and Signal Compression*. Kluwer Academic Press/Springer, 1992.

M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, pages 136–143, July 1963.

C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in Neural Information Processing Systems*, 24, 2011.

J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1), 2006.

A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76: 210–225, 2012.

B. Jedynak, P. I. Frazier, and R. Sznitman. Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability*, 49:114–136, 2012.

S. C. Jun, S. M. Plis, D. M. Ranken, and D. M. Schmidt. Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data. *Physics in Medicine and Biology*, 51: 5549–5564, 2006.

E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437471, 1996.

S. Kar and J. M. F. Moura. Covergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs. *IEEE Journal of Selected Topics in Signal Processing*, 5(4), August 2011.

N. El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, 36(6): 2757–2790, 2008.

R. M. Karp and R. Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 881–890, 2007.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press US, first edition, 1996.

Charles Van Loan and Nikos Pitsianis. Approximation with kronecker products. In *Linear Algebra for Large Scale and Real Time Applications*, pages 293–314. Kluwer Publications, 1993.

P. Molavi, A. Jadbabaie, K. R. Rad, and A. Tahbaz-Salehi. Reaching consensus with increasing information. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):358–369, April 2013.

M. Nokleby, W. Bajwa, R. Calderbank, and B. Aazhang. Toward resource-optimal consensus over the wireless medium. *IEEE Journal of Selected Topics in Signal Processing*, 7(2), April 2013.

R. J. Owen. A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70:351–356, 1975.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

N. Raj Rao, J. Mingo, R. Speicher, and A. Edelman. Statistical eigen-inference from large wishart matrices. *Annals of Statistics*, 36(6): 2850–2885, 2008.

A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2: 494–515, 2008.

A. Rucci, S. Tebaldini, and F. Rocca. Skp-shrinkage estimator for sar multi-baselines applications. In *Proceedings of IEEE Radar Conference*, 2010.

V. Saligrama, M. Alanyali, and O. Savas. Distributed detection in sensor networks with packet loss and finite capacity links. *IEEE Transactions on Signal Processing*, 54(11):4118–4132, November 2006.

S. Tebaldini. Algebraic synthesis of forest scenarios from multibaseline polinsar data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(12), December 2009.

T. Tsiligkaridis and A. O. Hero. Low separation rank covariance estimation using kronecker product expansions. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, July 2013a.

T. Tsiligkaridis and A. O. Hero. Covariance Estimation via Kronecker Product Expansions. *arXiv: 1302.2686*, February 2013b.

T. Tsiligkaridis, A. O. Hero, and S. Zhou. Convergence Properties of Kronecker Graphical Lasso algorithms. *arXiv:1204.0585*, July 2012.

T. Tsiligkaridis, A. O. Hero, and S. Zhou. On Convergence of Kronecker Graphical Lasso Algorithms. *IEEE Transactions on Signal Processing*, 61(7):1743–1755, April 2013a.

T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. Collaborative 20 Questions for Target Localization. *Preprint, arXiv: 1306.1922*, August 2013b.

T. Tsiligkaridis, B. M. Sadler, and A. O. Hero. A Collaborative 20 Questions model for target search with human-machine interaction. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013c.

J. Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachussets Institute of Technology, Cambridge, MA, November 1984.

J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, September 1986.

R. Waeber, P. I. Frazier, and S. G. Henderson. Bisection search with noisy responses. *SIAM Journal of Control and Optimization*, 53(3): 2261–2279, 2013.

H. Wainer. *Computerized Adaptive Testing: A Primer*. Routledge, 2 edition, 2000.

K. Werner, M. Jansson, and P. Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2), February 2008.

Karl Werner and Magnus Jansson. Estimation of kronecker structured channel covariances using training data. In *Proceedings of EUSIPCO*, 2007.

A. Wiesel, Y. Eldar, and A. O. Hero. Covariance estimation in decomposable gaussian graphical models. *IEEE Transactions on Signal Processing*, 58(3):1482–1492, March 2010.

J. Yin and H. Li. Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119–140, 2012.

K. Yu, J. Lafferty, S. Zhu, and Y. Gong. Large-scale collaborative prediction using a nonparametric random effects model. *ICML*, pages 1185–1192, 2009.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94:19–35, 2007.

Y. Zhang and J. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in Neural Information Processing Systems*, 23:2550–2558, 2010.