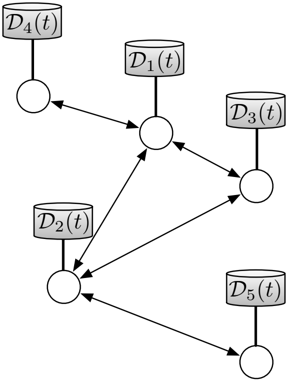
Network Science
Models and Distributed Algorithms

IST-CMU Phd course
João Xavier
TA: João Martins

November 9, 2016

Estimation in static undirected networks



- n agents; agent i holds **time-varying** dataset $\mathcal{D}_i(t)$
- the probability distribution of the $\mathcal{D}_i(t)$'s depends on parameter θ
- communication network is static and undirected
- communication happens in discrete time $t = 1, 2, 3, \ldots$
- goal: guess θ from $\bigcup_{s=1}^{t} \bigcup_{i=1}^{n} \mathcal{D}_i(s)$ at $t = 1, 2, 3, \ldots$

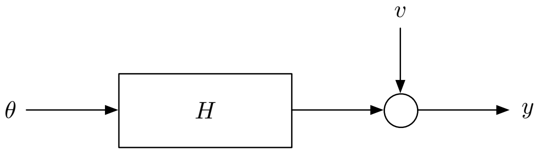
Crash course on parametric estimation

- a parametric estimation problem requires:
  - a parameter space Θ
  - an observation space Y
  - a family of probability density functions $\{p_\theta : Y \to \mathbf{R}\}_{\theta \in \Theta}$:

$$p_\theta(y) \geq 0 \text{ for all } y, \quad \int_Y p_\theta(y)dy = 1$$

- the estimation game:
  - mother Nature chooses a $\theta \in \Theta$ and uses p_θ to draw a sample y
  - you see y and have to guess θ

Example: noisy channel



- data model:
  - $\theta \in \mathbf{R}^p$ is the message
  - $H \in \mathbf{R}^{d \times p}$ is the channel, assumed full column-rank
  - $v \sim \mathcal{N}(0, \Sigma)$ is gaussian noise
  - $y \in \mathbf{R}^d$ is the measurement
  - goal is channel inversion: given y, guess θ

- corresponds to:
  - parameter space $\Theta = \mathbf{R}^p$
  - observation space $Y = \mathbf{R}^d$
  - parametric family

$$p_\theta(y) = \frac{1}{(2\pi)^{d/2} \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(y - H\theta)^T \Sigma^{-1} (y - H\theta)}$$

- an estimator is a map

$$\widehat{\theta} : Y \to \Theta, \qquad y \mapsto \widehat{\theta}(y)$$

(implemented as a closed-form expression, a matlab file, ...)

- an estimator is a **random** variable: it acts on the random variable y

- distribution of y varies with the θ chosen by mother Nature

- so, distribution of $\widehat{\theta}$ also varies with the θ chosen by mother Nature

- mean value of $\widehat{\theta}$ is

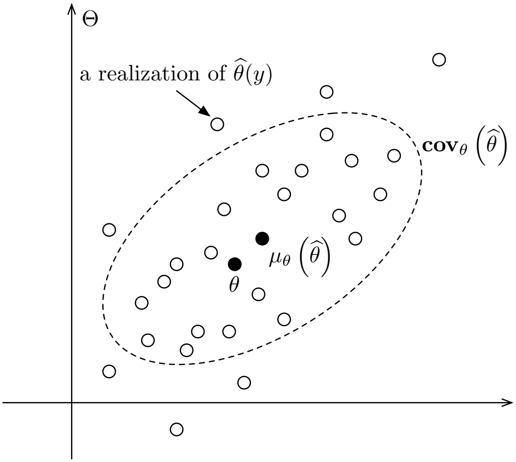
$$\begin{aligned}\mu_\theta\left(\widehat{\theta}\right) &= \mathbf{E}_\theta\left(\widehat{\theta}(y)\right)\\ &= \int_Y \widehat{\theta}(y)p_\theta(y)dy\end{aligned}$$

- interpretation: $\mu_\theta\left(\widehat{\theta}\right)$ is the mean value of $\widehat{\theta}$ when mother Nature chooses θ

- covariance of $\widehat{\theta}$ is

$$\begin{aligned}\textbf{cov}_\theta\left(\widehat{\theta}\right) &= \mathbf{E}_\theta\left(\left(\widehat{\theta}(y)-\mu_\theta\left(\widehat{\theta}\right)\right)\left(\widehat{\theta}(y)-\mu_\theta\left(\widehat{\theta}\right)\right)^T\right)\\ &= \int_Y \left(\widehat{\theta}(y)-\mu_\theta\left(\widehat{\theta}\right)\right)\left(\widehat{\theta}(y)-\mu_\theta\left(\widehat{\theta}\right)\right)^T p_\theta(y)dy\end{aligned}$$

- interpretation: $\textbf{cov}_\theta\left(\widehat{\theta}\right)$ tell us how $\widehat{\theta}$ spreads when mother Nature chooses θ



- what is a perfect estimator?

$$\mu_\theta\left(\widehat{\theta}\right) = \theta \qquad \mathbf{cov}_\theta\left(\widehat{\theta}\right) = 0 \quad \text{for } \textbf{all } \theta \in \Theta$$

- Cramér-Rao bound says perfect estimators cannot exist:
  - if $\widehat{\theta}$ is unbiased,
$$\mu_\theta\left(\widehat{\theta}\right) = \theta \quad \text{for all } \theta \in \Theta,$$
  - then
$$\mathbf{cov}_\theta\left(\widehat{\theta}\right) \succeq I(\theta)^{-1}$$
    where
$$I(\theta) = \mathbf{E}_\theta\left(-\nabla_\theta^2 \log p_\theta(y)\right)$$
    is the Fisher information matrix

- sometimes, we can design efficient estimators:
$$\mu_\theta\left(\widehat{\theta}\right) = \theta \qquad \mathbf{cov}_\theta\left(\widehat{\theta}\right) = I\left(\theta\right)^{-1} \quad \text{for } \textbf{all } \theta \in \Theta$$

- how can we design efficient estimators?

- sometimes, the maximum likelihood (ML) principle works:

$$\widehat{\theta}_{\mathsf{ML}} : Y \to \Theta, \qquad y \mapsto \arg\max_{\theta \in \Theta} p_\theta(y)$$

(finds the θ that makes the observation the most plausible)

- in general, solving the ML optimization problem is hard...

Example: noisy channel from page 4

- ML estimator is

$$\widehat{\theta}_{\mathsf{ML}}(y) = \left(H^T \Sigma^{-1} H\right)^{-1} H^T \Sigma^{-1} y$$

- $\widehat{\theta}_{\mathsf{ML}}$ is unbiased:

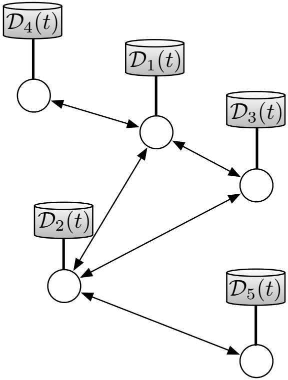
$$\mu_\theta \left(\widehat{\theta}_{\mathsf{M}L}\right) = \theta \quad \text{for all } \theta \in \Theta$$

- the covariance of $\widehat{\theta}_{\mathsf{ML}}$ is

$$\mathbf{cov}_\theta \left(\widehat{\theta}_{\mathsf{ML}}\right) = \left(H^T \Sigma^{-1} H\right)^{-1}$$

- $\widehat{\theta}_{\mathsf{ML}}$ is efficient because $I(\theta) = H^T \Sigma^{-1} H$ for all $\theta \in \Theta$

Distributed parameter estimation



- at time t, agent i observes $y_i(t) = H_i(t)\theta + v_i(t)$: $\mathcal{D}_i(t) = \{y_i(t)\}$
- $v_i(t) \sim \mathcal{N}(0, \sigma^2 I)$ is independent across agents and time
- agent i only knows its measurements $y_i(t)$ and matrices $H_i(t)$
- goal: guess θ from $\bigcup_{s=1}^{t} \bigcup_{i=1}^{n} \mathcal{D}_i(s)$ at $t = 1, 2, 3, \ldots$

- what would a centralized estimator do?

- at time t, a centralized estimator would know:
  - $H_i(s)$ for all i and $s = 1, 2, \ldots, t$ (all the sensing matrices up to t)
  - $\bigcup_{s=1}^{t} \bigcup_{i=1}^{n} y_i(s)$ (all the network observations up to t)

- data model from the perspective of the central node, at time t:

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(t) \end{bmatrix} = \underbrace{\begin{bmatrix} H(1) \\ H(2) \\ \vdots \\ H(t) \end{bmatrix}}_{\mathcal{H}(t)} \theta + \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(t) \end{bmatrix}$$

where

$$y(t) := \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} \quad H(t) := \begin{bmatrix} H_1(t) \\ H_2(t) \\ \vdots \\ H_n(t) \end{bmatrix} \quad v(t) := \begin{bmatrix} v_1(t) \\ v_2(t) \\ \vdots \\ v_n(t) \end{bmatrix}$$

- assumptions:
  - (bounded sensing) $\|H_i(t)\| \leq M$, for all i and t
  - (unbounded information) $\mathcal{H}(t)^T \mathcal{H}(t) \to \infty$ as $t \to \infty$

- at time t, the central node could implement the ML estimator:

$$\widehat{\theta}_{\mathsf{ML}}(t) = \Bigg(\underbrace{\frac{1}{nt} \sum_{s=1}^{t} H(s)^T H(s)}_{P(t)} \Bigg)^{-1} \Bigg(\underbrace{\frac{1}{nt} \sum_{s=1}^{t} H(s)^T y(s)}_{z(t)} \Bigg)$$

- the inverse of $P(t)$ exists for large t because $\mathcal{H}(t)^T \mathcal{H}(t) \to \infty$

- note the recursions:

$$\begin{aligned} z(t+1) &= \frac{t}{t+1} z(t) + \frac{1}{t+1} \left(\frac{1}{n} \sum_{i=1}^{n} H_i(t+1)^T y_i(t+1) \right) \\ P(t+1) &= \frac{t}{t+1} P(t) + \frac{1}{t+1} \left(\frac{1}{n} \sum_{i=1}^{n} H_i(t+1)^T H_i(t+1) \right) \end{aligned}$$

- ML estimator is unbiased and has covariance

$$\begin{aligned}\textbf{cov}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right) &= \sigma^2\left(\sum_{s=1}^{t} H(s)^T H(s)\right)^{-1} \\ &= \sigma^2\left(\mathcal{H}(t)^T\mathcal{H}(t)\right)^{-1}\end{aligned}$$

- we have

$$\textbf{cov}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right) \to 0 \quad \text{as } t \to \infty$$

  because $\mathcal{H}(t)^T\mathcal{H}(t) \to \infty$

- the ML estimator gets more and more accurate as time goes: at $t = \infty$ it "knows" θ exactly

- structure of ML estimator suggests the estimator at agent i:

$$\widehat{\theta}_i(t) = P_i(t)^{-1} z_i(t)$$

where $z_i(t)$ and $P_i(t)$ are local estimates of $z(t)$ and $P(t)$

- agent i updates $P_i(t)$ and $z_i(t)$ as follows:

$$\begin{aligned} z_i(t+1) &= \frac{t}{t+1} \sum_{j \sim i} W_{ij} z_j(t) + \frac{1}{t+1} H_i(t+1)^T y_i(t+1) \\ P_i(t+1) &= \frac{t}{t+1} \sum_{j \sim i} W_{ij} P_j(t) + \frac{1}{t+1} H_i(t+1)^T H_i(t+1) \end{aligned}$$

- $W \in \mathbf{R}^{n \times n}$ is a primitive matrix with diagonal entries, $W\mathbf{1} = \mathbf{1}$ and $W_{ij} = 0$ if $i \not\sim j$

- how good is the distributed estimator $\widehat{\theta}_i$?

- each $\widehat{\theta}_i(t)$ is unbiased:

$$\mathbf{E}_\theta\left(\widehat{\theta}_i(t)\right) = \theta \quad \text{for all } t \text{ and } \theta \in \Theta$$

- each $\widehat{\theta}_i(t)$ is asymptotically equivalent to $\widehat{\theta}_{\mathsf{ML}}(t)$:

$$\Sigma_\theta(t)^{-\frac{1}{2}}\Upsilon_\theta(t)\Sigma_\theta(t)^{-\frac{1}{2}} \to I, \quad \text{as } t \to \infty,$$

  for all $\theta \in \Theta$, where

$$\Sigma_\theta(t) := \mathbf{cov}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right) \quad \text{and} \quad \Upsilon_\theta(t) := \mathbf{cov}_\theta\left(\widehat{\theta}_i(t)\right)$$

Proof for scalar parameter θ

- data model at agent i:

$$y_i(t) = h_i(t)\theta + v_i(t)$$

  where

  - $\theta \in \mathbf{R}$
  - $h_i(t) \in \mathbf{R}$
  - $v_i(t) \sim \mathcal{N}(0, \sigma^2)$

- in vector notation, the network measurement at time t is

$$y(t) = h(t)\theta + v(t)$$

  where

$$y(t) := \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix} \quad h(t) := \begin{bmatrix} h_1(t) \\ h_2(t) \\ \vdots \\ h_n(t) \end{bmatrix} \quad v(t) := \begin{bmatrix} v_1(t) \\ v_2(t) \\ \vdots \\ v_n(t) \end{bmatrix}$$

- data model from the perspective of the central node, at time t:

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(t) \end{bmatrix} = \begin{bmatrix} h(1) \\ h(2) \\ \vdots \\ h(t) \end{bmatrix} \theta + \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(t) \end{bmatrix}$$

- ML estimator is

$$\widehat{\theta}_{\mathsf{ML}}(t) = \frac{\frac{1}{nt} \sum_{s=1}^{t} h(s)^T y(s)}{\frac{1}{nt} \sum_{s=1}^{t} \|h(s)\|^2}$$

- ML estimator is unbiased and has variance

$$\mathbf{var}_\theta \left(\widehat{\theta}_{\mathsf{ML}}(t) \right) = \frac{\sigma^2}{\sum_{s=1}^{t} \|h(s)\|^2}$$

- estimator at agent i is

$$\widehat{\theta}_i(t) = \frac{z_i(t)}{p_i(t)}$$

where

$$\begin{aligned} z_i(t+1) &= \frac{t}{t+1}\sum_{j\sim i} W_{ij} z_j(t) + \frac{1}{t+1} h_i(t+1) y_i(t+1) \\ p_i(t+1) &= \frac{t}{t+1}\sum_{j\sim i} W_{ij} p_j(t) + \frac{1}{t+1} h_i(t+1)^2 \end{aligned}$$

- in vector notation:

$$z(t+1) = \frac{t}{t+1} W z(t) + \frac{1}{t+1} h(t+1) \odot y(t+1) \quad (1)$$

$$p(t+1) = \frac{t}{t+1} W p(t) + \frac{1}{t+1} h(t+1) \odot h(t+1) \quad (2)$$

- taking expected values in (1):

$$\mathbf{E}_\theta\left(z(t+1)\right) = \frac{t}{t+1} W \mathbf{E}_\theta\left(z(t)\right) + \frac{1}{t+1} h(t+1) \odot h(t+1)\theta$$

- so,

$$\mathbf{E}_\theta\left(z(t)\right) = p(t)\theta \quad \text{for all } t$$

- we conclude

$$\begin{aligned} \mathbf{E}_\theta\left(\widehat{\theta}_i(t)\right) &= \frac{\mathbf{E}_\theta\left(z_i(t)\right)}{p_i(t)} \\ &= \theta \quad \text{for all } \theta \in \Theta \end{aligned}$$

(i.e., each $\widehat{\theta}_i(t)$ is unbiased)

- the variance of $\widehat{\theta}_i(t)$ is

$$\textbf{var}_\theta\left(\widehat{\theta}_i(t)\right) = \frac{\textbf{var}_\theta(z_i(t))}{p_i(t)^2} = \frac{\mathbf{E}_\theta\left(z_i(t) - p_i(t)\theta\right)^2}{p_i(t)^2}$$

- let $\delta_i(t) := z_i(t) - p_i(t)\theta$ and

$$\delta(t) := \begin{bmatrix} \delta_1(t) \\ \delta_2(t) \\ \vdots \\ \delta_n(t) \end{bmatrix} = z(t) - p(t)\theta$$

- we have the recursion

$$\delta(t+1) = \frac{t}{t+1} W \delta(t) + \frac{1}{t+1} h(t+1) \odot v(t+1)$$

- decomposing $\delta(t) = \overline{\delta}(t)\mathbf{1} + U\widehat{\delta}(t)$ we have

$$\overline{\delta}(t+1) = \frac{t}{t+1}\overline{\delta}(t) + \frac{1}{n(t+1)}h(t+1)^T v(t+1) \qquad (3)$$

$$\widehat{\delta}(t+1) = \frac{t}{t+1}\Lambda\widehat{\delta}(t) + \frac{1}{t+1}U^T h(t+1) \odot v(t+1) \qquad (4)$$

- equation (3) tells us

$$\overline{\delta}(t) = \frac{1}{nt}\sum_{s=1}^{t} h(s)^T v(s) \qquad \text{for all } t \geq 1$$

- so,

$$\mathbf{var}_\theta\left(t\overline{\delta}(t)\right) = \frac{\sigma^2}{n^2}\sum_{s=1}^{t} \|h(s)\|^2$$

- equation (4) tell us that $\textbf{var}_\theta \left(t\widehat{\delta}(t) \right)$ is bounded

- from equation (2) we have

$$\begin{aligned} \overline{p}(t+1) &= \frac{t}{t+1}\overline{p}(t) + \frac{1}{n(t+1)} \left\| h(t+1) \right\|^2 \\ \widehat{p}(t+1) &= \frac{t}{t+1}\Lambda\widehat{p}(t) + \frac{1}{t+1}U^T h(t+1) \odot h(t+1) \end{aligned}$$

- it follows that

$$t\overline{p}(t) = \frac{1}{n}\sum_{s=1}^{t} \left\| h(s) \right\|^2$$

and $t\widehat{p}(t)$ is bounded

- we have

$$\frac{\mathbf{var}_\theta\left(\widehat{\theta}_i(t)\right)}{\mathbf{var}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right)} = \frac{\mathbf{var}_\theta\left(\delta_i(t)\right)}{p_i(t)^2\mathbf{var}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right)} = \frac{\mathbf{var}_\theta\left(\delta_i(t)\right)}{p_i(t)}\frac{1}{p_i(t)\mathbf{var}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right)}$$

- there holds:

$$\frac{t\mathbf{var}_\theta\left(\delta_i(t)\right)}{p_i(t)} \to \frac{\sigma^2}{n} \quad \text{and} \quad \frac{1}{tp_i(t)\mathbf{var}_\theta\left(\widehat{\theta}_{\mathsf{ML}}(t)\right)} \to \frac{n}{\sigma^2}$$

- if $\{A(t)\}_{t\geq 0}, \{B(t)\}_{t\geq 0}$ are sequences of positive-definite matrices and

$$A^{-1/2}(t)B(t)A^{-1/2}(t) \to I,$$

then

$$\frac{\mathbf{tr}\,(B(t))}{\mathbf{tr}\,(A(t))} \to 1.$$

Useful formulas for random vectors

- if $x \in \mathbf{R}$ is a random variable and $f : \mathbf{R} \to \mathbf{R}$ a function,

$$\mathbf{E}(f(x)) = \int_{\mathbf{R}} f(x)p(x)dx,$$

where $p : \mathbf{R} \to \mathbf{R}$ is the probability density function of x

- if $X = (x_{ij}) \in \mathbf{R}^{n\times m}$ is a random matrix,

$$\mathbf{E}(X) = \begin{bmatrix} \mathbf{E}(x_{11}) & \mathbf{E}(x_{12}) & \cdots & \mathbf{E}(x_{1m}) \\ \mathbf{E}(x_{21}) & \mathbf{E}(x_{22}) & \cdots & \mathbf{E}(x_{2m}) \\ \vdots & & & \\ \mathbf{E}(x_{n1}) & \mathbf{E}(x_{n2}) & \cdots & \mathbf{E}(x_{nm}) \end{bmatrix}$$

(random vectors correspond to $m = 1$)

- the covariance of a random vector $x \in \mathbf{R}^n$ is

$$\mathbf{cov}\,(x) = \mathbf{E}\left((x - \mathbf{E}(x))\,(x - \mathbf{E}(x))^T\right) \in \mathbf{R}^{n \times n}$$

- the variance of a random vector $x \in \mathbf{R}^n$ is

$$\mathbf{var}\,(x) = \mathbf{tr}\,(\mathbf{cov}\,(x))$$

- note that

$$\mathbf{var}(x) = \mathbf{E}\left(\|x - \mathbf{E}(x)\|^2\right)$$

- the covariance between random vectors $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$ is

$$\mathbf{cov}(x, y) = \mathbf{E}\left((x - \mathbf{E}(x))\,(y - \mathbf{E}(y))^T\right) \in \mathbf{R}^{n \times m}$$

(note: we use $\mathbf{cov}(x) \equiv \mathbf{cov}(x, x)$)

- we say that the random vectors x and y are uncorrelated if

$$\mathbf{cov}(x, y) = 0$$

- if x is a random vector,

$$\mathbf{E}\,(Ax) = A\mathbf{E}(x) \qquad \mathbf{cov}(Ax) = A\mathbf{cov}(x)A^T$$

- if x and y are random vectors,

$$\begin{aligned}\mathbf{E}(x+y) &= \mathbf{E}(x)+\mathbf{E}(y)\\ \mathbf{cov}(x+y) &= \mathbf{cov}(x)+\mathbf{cov}(y)+\mathbf{cov}(x,y)+\mathbf{cov}(y,x)\end{aligned}$$

- x and y are **independent** random vectors if

$$p(x,y) = p(x)p(y)$$

(joint pdf is the product of the marginal pdfs)

- if x and y are independent random vectors,

$$\mathbf{cov}(x+y) = \mathbf{cov}(x)+\mathbf{cov}(y)$$

- if (x,y) is a gaussian random vector and x and y are uncorrelated, then x and y are independent

- if x and y are random vectors such that $x \leq y$, then

$$\mathbf{E}(x) \leq \mathbf{E}(y)$$

To know more

- distributed estimation:
  - Z. Weng and P. Djurić, "Efficient estimation of linear parameters from correlated node measurements over networks," *IEEE Sig. Proc. Lett.*, 21(11), 2014.

- background on statistical signal processing:
  - S. Kay, *Fundamentals of Statistical Signal Processing, vol 1: Estimation Theory*, Prentice Hall.
  - L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*.