10-708 modelling networks 03/06/2020

L8 - Gaussian graphical models and Ising models

- Practical problem in real-world → new algorithmic design
(*) Network research has been active research area for a long time.
(*) ML/DL → default image/text data
(*) network research deals with relationships.
(*) EX: where does the graph come from?
  - Given, or we need to discover them?
(*) Jesus network → how is graph constructed?
  - Cooccurrences in same use, chapter, ?
  - A source of variation
  - What is considered a natural object can be a man-made artefact.
(*) EX trying to initiate a debate on
  i) the ontological status of a graph
  ii) interrogating its truth/representation value.
(*) EX: None of the common networks are physical, 'undisputable' facts.

(*) Evolving networks
- friends and foes are interchangeable (e.g. politics)

(*) ML structural learning for completely observed GMs
- EX: Previously assumed structure was given.
- Very complex topic fraught with mathematical and algorithmic challenges.
- Trail has gone colder due to low-hanging fruit in other areas.
(*) (A1): H/W: go through Chow-Liu algorithm; estimate graph structure
- examine a number of 'optimal' algorithms; focus on
- Pairwise MRFs: covariance selection, neighbourhood selection.

(*) chow-liu gives an 'optimal tree'

- Ex: focus on arbitrary graph topology today

(*) Network inference as parameter estimation

- we have a model for parametrising social net GM

Ex: what is this called?

W: Feature based MGM? × Pairwise MRF / Boltzmann Machine

Ex: tree is a tight connection between parameters and structure

Ex: Turn a graph structure inference problem ⟶ parameter estimation problem
(use a placeholder?)

- some θs may be non-zero, some zero → indicates presence of edges

(*) Distinct from chow-liu ⟶ C-L scores edges.

(*) Pairwise MRFs

- weights defined over singleton and pairwise potentials only

(on):- $p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_{12} x_1 x_2 + \theta_{13} x_1 x_3 + \theta_{23} x_2 x_3 + \theta_{34} x_3 x_4\}$

- Nodal states discrete → Ising/Potts model

cont. → Gaussian graphical models (WHY?)

(*) positional location of parameters in parameter matrix encodes structure

- Rewrite MVG as a continuous MRF (*) write out
- WOLG let $\mu = 0$ $Q = \Sigma^{-1}$

$$p(x_1, x_2, x_3, \ldots, x_p \mid \mu = 0, Q) = \frac{|Q|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2} \sum_i q_{ii} (x_i)^2 - \sum_{i<j} q_{ij} x_i x_j\right\}$$

①

- off-diagonal elements in Q - weights of pairwise potentials
- on diagonal ——— 〃 ——— — 〃 — singleton — 〃 —

(*) Covariance estimation → precision matrix estimation?
(*) Continuous observations → estimate parameters of Gaussian?
Ex: will this work?

(*) Gaussian Graphical Model.

$$\underline{x}^{(n)} \sim N(\underline{0}, \underline{\underline{\Sigma}}^{(n)})$$

$\underline{x}^{(n)}$: e.g. microarray samples, all types
$\underline{\underline{\Sigma}}^{(n)}$: encodes dependencies among genes

(*) Precision matrix

$$\underline{\underline{\Omega}}^{(n)} = (\underline{\underline{\Sigma}}^{(n)})^{-1}$$ - encodes non-zero edges in GGM.

(*) Markov vs Correlation Network
Ex: Previously when ppl. dealt with networks, precision matrices used to define a correlation network (a way of finding structures).

(*) correlation network → covariance matrix (1) - A design choice
(*) GGM - Markov network → precision matrix. (2)

Ex: we do not use (1) as it is not consistent with our definitions of conditional independencies.
Ex: MN more meaningful network structure, as we don't often look at a pair of variables (in a graph) in isolation from context (i.e. not given other variables)
- 2 variables independent given the other variables

1) correlation network:-

$$\Sigma_{i,j} = 0 \Rightarrow X_i \perp X_j \text{ or } p(X_i, X_j) = p(X_i)\,p(X_j)$$

(?? x)

2) C.I / pCC offer better dep measure

$$\Omega_{i,j} = 0 \Rightarrow X_i \perp X_j \mid \underline{X}_{-ij} \text{ or } p(X_i, X_j \mid \underline{X}_{-ij}) = p(X_i \mid \underline{X}_{-ij})\,p(X_j \mid \underline{X}_{-ij})$$

(*) Sparsity

- later

(*) Network structure learning with LASSO

(#2): Argument overall here

- EX: Is there a false distinction between CN and MN?
  - After all, precision matrix and covariance matrix are mathematically related via inversion

↓

W: Depends on matrix properties governing possibility of inversion (e.g. singularity, positive definiteness, rank, eigenvalues) ✓

(*) Assuming covariance invertible introduces hidden assumptions
- full rank covariance matrix
- rank deficiency ⇒ non invertible covariance ⇏ negation of MN as a model.

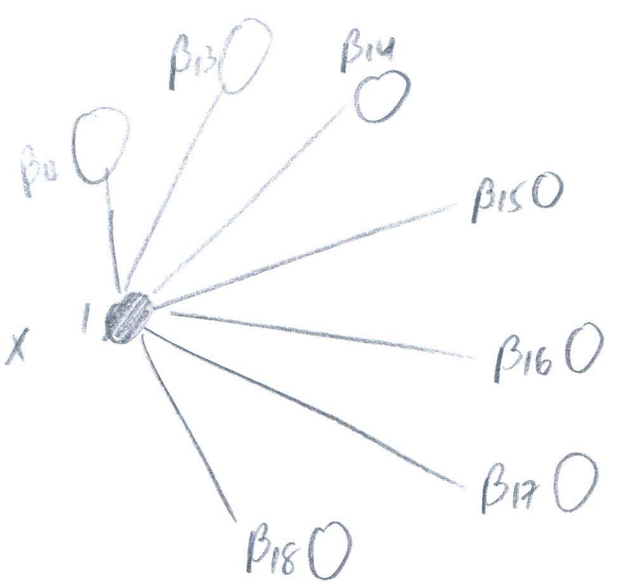
(*) Computational issues from matrix inversion $O(N^3)$

↓

EX: Estimation of Θ can be done piece-wise algorithmically.

(*) Nodely algorithms, learn structure incrementally.

(try this)



$$X_i = \beta X_i + \epsilon$$

- LASSO regression of all nodes to a target node.
- Assume network is a GGM

- estimate if there is an edge between 2 nodes or not.
- one-node neighbourhood problem.

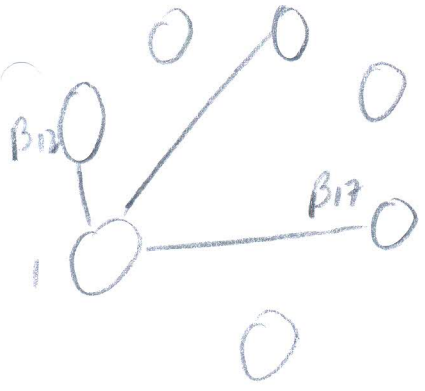
EX: Does this yield a MN equivalent?

(*) Start with
- neighbourhood of every node, examine each node one by one and use regression approach to det. if edges are zero.

Ex: recouple questions
- cs/engineering approach (more track-and-build)
  - ↳ does algorithm yield a graph?
  - ↳ is the graph 'correct'?

(*) operational problem in above thought experiment.
- regression will not yield non-zero β_i
- All nodes will be your neighbour → fully connected graph
- so how to amend to get a meaningful graph

(*) Add sparse regularisation (L_1)
- $\hat{\beta}_1 = \text{argmin}_{\beta_1} \|Y - X\beta_1\|_2^2 + \lambda\|\beta_1\|_1$



(*) L_1 regularisation
- sparsity enforcing constraint via L_1 norm (diamond)
  L_2 norm (circle) (no sparsity)
  L_p ...

(*) Recall why L_1 norm → sparsity

(*) L_0 norm - spikes → non-convex; difficulty using gradients.

- Find/solve using Lagrangians
- Also known as neighbourhood selection

(*) Tune will be some edges non-zero, rest pushed to zero (sparsity)

(*) Network learning with LASSO ← introduce a 'rule' to encode intuition (A1)
- Repeat for every node
- Form total edge set $\hat{E} = \{(u,v) : \max(|\hat{\beta}_{uv}|, |\hat{\beta}_{vu}|) > 0\}$

(*) so we now have a meaningful graph

(*) we have 2 estimates of every edge due to mutuality of neighbours; one as target; one as input. ✓

(*) Here we introduce above rule
- edge features if one of the regression coefficient estimates is non-zero. (i.e. whether edge present/relevance ⇒ C.I.

Ex: whether this a MRF consistent with previous definition → a concern about validity of graph based on definition

Ex: neighbourhood selection is extremely simple computational complexity wise.

(*) Consistent structure Recovery (2006)

Ex: Meinshausen & Bühlmann

- under certain conditions; graph estimated by neighbourhood selection/ LASSO is the same graph described in GGM wrt to precision matrix of GGM.

- for continuous r.v.

(TT): If $\lambda_S > C\sqrt{\frac{\log p}{s}}$ then with high prob.

$S(\hat{\beta}) \longrightarrow S(\beta^*)$

(*) Read papers structure disc.

(*) Provide intuition on why algorithm works

i.e. repeated application of neighbourhood selection yields (LASSO reg)
pairwise graphical model consistent with GGM.

(*) MVG:

- some interesting properties.

- use some partitioning

(w): Remember Bishop/Murphy → conditional, marginal Gaussians + matrix inversion lemma.

- $p\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mid \mu, \Sigma\right) = N\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mid \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$ - Joint Gaussian over subsections indexed by part.

(?): How to write down $p(x_2)$, $p(x_1|x_2)$, $p(x_2|x_1)$?

- must satisfies theoretical properties

$p(x_2) = N(x_2 | \mu_2^m, V_2^m)$ $\quad p(x_1|x_2) = N(x_1 | \mu_{12}, V_{12})$

$\mu_2^m = \mu_2$

$V_2^m = \Sigma_{22}$

$\mu_{12} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$

$V_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

adjusted μ_1 and Σ_{11} with offset terms.

- use a lot when dealing with linear systems.

EX: commit to memory

(*) Matrix-inversion lemma

- (A3) - Review ✓

(*) covariance and precision matrices.

- Establish relation and properties between covariance and precision in context of Gaussians ✓

- single node conditional

- Can use these results to write down conditional distr of one node given all other nodes IF the whole distr is Gaussian

(*) $p(x_i | x_{-i})$

- Will let $\mu = 0$

- $p(x_i | x_{-i})$ is Gaussian with mean and covariance given in slides.

(*) $p(X_i | \underline{X}_{-i}) = N\left(\frac{\underline{q}_i^T}{-q_{ii}} \underline{X}_{-i}, q_{ii|-i}\right)$

- Note the q terms !

(*) EX: If you see conditional distri with Gaussian mean and covariance?

conditional auto-reg.

$X_i = \underline{\theta} \underline{X}_{-i} + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \Rightarrow X_i = \frac{\underline{q}_i^T}{-q_{ii}} \underline{X}_{-i} + \epsilon$

(*) neighbourhood selection → estimate θ → use rule to pose whether non-zero (we care only about structure)

- $p(X_i | \underline{X}_{-i}) = N\left(\frac{\underline{q}_i^T}{-q_{ii}} \underline{X}_{-i}, q_{ii|-i}\right)$

$Q = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}$ ← encodes MN of GGM

$q_{i|-i}$

$\frac{\underline{q}_i^T}{-q_{ii}} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$

$\hat{\underline{\theta}} = \begin{bmatrix} 0 \\ \hat{\theta}_0 \\ \vdots \\ \hat{\theta}_1 \\ 0 \\ 0 \end{bmatrix}$

(!!) sparseness of θ is the same as sparseness of Q ✓ (i.e. forget values only focus on whether component in $\underline{\theta}$ is zero (⇒ Q comp is zero)
(one-to-one correspondence?)

EX: If our goal is to estimate sparseness of $\underline{\theta}$ and Q, i.e. the parents locations that encode structure, then this achieves goal.

$S_i \equiv \{j : j \neq i, \theta_{ij} \neq 0\}$ (!!) = (?) ✓

EX: The maths is not the key message, rather insights / angles of viewing.

(!!) - Do it for all X_i $i = 1, \ldots, M$
- to get $\hat{Q}$ (also symmetric)

- sloppy

(*) Each estimation $\frac{\underline{q}_i^T}{q_{ii}}$ (via regression of X_i on $\underline{X}_{-i}$) yields an LASSO. estimate of a column of precision matrix

---

(*) Meinshausen-Bühlmann algorithm
(*) L1-regularised ML learning
(*) ...

} skipped

(*) learning Ising models (pairwise MRF)

- Assuming discrete valued nodes
- can we turn a structure learning problem into a param estim problem?

(1) $\underline{X} \to S_\lambda \to S \equiv Q \to \Sigma^{-1} \to N(\mu, \Sigma^{-1}) \to p(X_i | X_{\setminus i}) \to$ regression

- $\underline{X}$: data set of r.v.s.
- S_λ: 'structure' of graph
- $S \equiv Q$: turn estimat. of structure into est. of precision (structure est → param est.)
- Σ^{-1}: Gaussian - invoke it
- $N(\mu, \Sigma^{-1})$: (revisiting moves/traj of earlier lecture)

- some of these moves heuristic, some mathematical.

$$P(\underline{x} | \underline{\theta}) = \exp\left(\sum_{i \in V} \theta_{ii}^{\tau} x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j - A(\underline{\theta}) \right)$$

- not Gaussian due to $A(\theta)$, comput. of normaliser

Ex: How to deal with discrete situation? r.v.s.

(1) → empirical distri (i.e. counting)?

Ex: Hint: Invoke linear regression for continuous case as a placeholder for problem → alter placeholder

- modify to do a regression on discrete r.v.s. → logistic regression

(*) Build this kind of insight (heuristic)

$$p(x_{\setminus R} | x_{\setminus R}) = \text{logistic}\left(2 x_R \langle \theta_{\setminus R}, x_{\setminus R} \rangle \right)$$ with L1-regularisation.

(*) Proof → KKT / complementarity etc.

(*) scalar → vector valued.

- Ex: what is hidden assumption when moving from scalar to vector r.v.s.?
  - IID validity - IID ov scalars vs ov vectors

Ex: on something else problematises reg. of vector against scalar.

- modify to allow vector valued output. (Bishop)
- LINEAR/LOGISTIC → scalr output

(*) vector valued nodes

- ① How to deal with vectors?
- ② How to deal with vectors of variable length? → kernels on structured data (?)

(int) common in real world.

(W):

Ex: possible to directly estimate partial corelations of vectors of different lengths etc.

Q: How to estimate evolving networks?

- estimate a network from one bucket of votes at time t?
- statistical estimates need to be built on redundancy of data?
- You have only one data point
- IID is violated

(*) some issues for inferring structe of evolving networks.

(*) interest

- (Song, Kolar, Xing, 2009) - KELLER
- 1 data point, IPGM. - no possib of regression

(*) Premise is that one time step away network comes from same network but with deviations from baseline

- Kernel weighted L1 reg. logistic.

(*) other data points weighted by the distances from data point

(*) Still find a way of using the T data points.

(*) very interesting → find a way to use T data points for est.
systematic of time-evolving networks.

Ex: use tricks, that are tractable; and yield proof opportunities
on explicit

- diffcat to NN

(*) Inference II - skipped

- Read literature on how heuristic + theory work together.

(*) Rest of slides skipped.

- state of art on learning network structure

(SM): consistent, mathematically provable structure.

10-708 06/06/2020

L9 - CGMS, Ising models review

(*) matrix-inverse lemma

- for a block-partitioned matrix $\underline{M} = \begin{pmatrix} \underline{E} & \underline{F} \\ \underline{G} & \underline{H} \end{pmatrix}$ with $\underline{E}$ and $\underline{H}$ invertible.

- matrix-inverse lemma:-

$$(\underline{E} - \underline{F}\underline{H}^{-1}\underline{G})^{-1} = \underline{E}^{-1} + \underline{E}^{-1}\underline{F}(\underline{H} - \underline{G}\underline{E}^{-1}\underline{F})^{-1}\underline{G}\underline{E}^{-1}$$

method:

- diagonalise $\underline{M}$:-

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} E & F \\ G & H \end{bmatrix}\begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}$$

- Schur's complement: (notable in numerical analysis)

$$M/H = E - FH^{-1}G$$

- Note: for $XYZ = W$, $Y^{-1} = ZW^{-1}X$

$$M^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}\begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}^{-1}\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix}\begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix}\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & 0 \\ -H^{-1}G(M/H)^{-1} & H^{-1} \end{bmatrix}\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1}G(M/H)^{-1}FH^{-1} + H^{-1} \end{bmatrix}$$

- ~~Applying M.I.L:~~

$$\underline{\underline{M}}^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix} \quad (I)$$

- See Murphy (2012) 5.4.3.4.1
- (I) is achieved by decomposing M terms of $\underline{\underline{E}}$ and

$$M/E = (H - GE^{-1}F)$$

- We get MIL as a corollary

(*) Covariance and precision matrices

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \sigma_{11} & \underline{\sigma}_1^T \\ \underline{\sigma}_1 & \underline{\underline{\Sigma}}_{-1} \end{bmatrix}$$

σ_{11} - scalar
$\underline{\sigma}_1, \underline{\sigma}_1^T \in \mathbb{R}^{N-1}$
$\underline{\underline{\Sigma}}_{-1} \in \mathbb{R}^{(N-1)\times(N-1)}$

$$\underline{X} = \begin{bmatrix} X_i \\ | \\ \underline{X}_{-i} \\ | \end{bmatrix} \quad \underline{X} \in \mathbb{R}^N$$

$$\underline{\underline{M}}^{-1} = \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} \quad - \underline{\underline{Q}} = \underline{\underline{\Sigma}}^{-1}$$

$$(M/H)^{-1} = \underline{\sigma}_{11}^{-1} + \underline{\sigma}_{11}^{-1}\underline{\sigma}_1^T(\underline{\underline{\Sigma}}_{-1} - \underline{\sigma}'\underline{\sigma}_{11}^{-1}\underline{\sigma}_1^T)^{-1}\underline{\sigma}_1\underline{\sigma}_{11} = q_{11} \quad \checkmark$$

$$\underline{\underline{Q}} = \begin{bmatrix} q_{11} & -q_{11}\underline{\sigma}_1^T\underline{\underline{\Sigma}}_{-1}^{-1} \\ -q_{11}\underline{\underline{\Sigma}}^{-1}\underline{\sigma}_1 & \underline{\underline{\Sigma}}_{-1}^{-1}(I + q_{11}\underline{\sigma}_1\underline{\sigma}_1^T\underline{\underline{\Sigma}}_{-1}^{-1}) \end{bmatrix} = \begin{bmatrix} q_{11} & \underline{q}_1^T \\ \underline{q}_1 & \underline{\underline{Q}}_{-1} \end{bmatrix}$$

(*) Single node conditional

- Partition $\underline{X} = \begin{bmatrix} X_i \\ \underline{X}_{-i} \end{bmatrix}$ $\quad p\left(\begin{bmatrix} X_i \\ \underline{X}_{-i} \end{bmatrix} \middle| \underline{\mu}, \underline{\underline{\Sigma}}\right) = \mathcal{N}\left(\begin{bmatrix} X_i \\ \underline{X}_{-i} \end{bmatrix} \middle| \begin{bmatrix} \underline{\mu}_i \\ \underline{\mu}_{\underline{X}_{-i}} \end{bmatrix}, \begin{bmatrix} \underline{\underline{\Sigma}}_{X_iX_i} & \underline{\underline{\Sigma}}_{X_i\underline{X}_{-i}} \\ \underline{\underline{\Sigma}}_{\underline{X}_{-i}X_i} & \underline{\underline{\Sigma}}_{\underline{X}_{-i}\underline{X}_{-i}} \end{bmatrix}\right)$