

18- Gaussian graphical models and Ising models

- practical problem in real-world \rightarrow new algorithmic design
- (*) Network research has been active research area for a long time.
- (*) ML/DL \rightarrow 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 \rightarrow how is graph constructed?
 - co-occurrences in same user, 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.

19-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.
- field has gone colder due to low-hanging fruit in other areas.
- (*) Q1: HW: 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 uGM? x pairwise MRF/Boltzmann Machine

ex: tree is a tight connection between parameters and structure

ex: turn a graph structure inference problem \rightarrow parameter estimation problem
(use a placeholder?)

- some θ_i s may be non-zero; some zero \rightarrow indicates presence of edges

(*) distinct from Chow-Liu \rightarrow C-L scores edges.

(*) Pairwise MRFs

- weights defined our singletor and pairwise potentials only

$$\text{defn: } p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp \left\{ \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_{12} x_1 x_2 + \theta_{13} x_1 x_3 + \theta_{14} x_1 x_4 + \theta_{23} x_2 x_3 + \theta_{24} x_2 x_4 + \theta_{34} x_3 x_4 \right\}$$

- Nodal states discrete \rightarrow Ising/Wolff model

cont. \rightarrow Gaussian graphical models (HTY?)

(*) Positional location of parameters in param matrix encodes structure

- Rewrite MVG as a continuous MRF $\quad \text{(*) write out}$

- WOLG let $\mu=0, Q=\Sigma^{-1}$

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

- off-diagonal elements in Q - weights of pairwise potentials

- on diagonal $\text{---} \parallel \text{---} \quad \text{---} \parallel \text{---}$ singletor $\text{---} \parallel \text{---}$

- (*) covered weight estimation over potentials \rightarrow precision matrix estimation?
 (*) continuous observations \rightarrow estimate form of Gaussian?
 ex: will this work?

(*) Gaussian Graphical Model.

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

↑ ↑
 e.g. microarray samples encodes dependencies
 all types among genes

(*) Precision matrix

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

(*) Markov vs Correlation Network

Previously when ppl. dealt with networks; precision matrices used to define a correlation network (a way of finding structures).

(*) correlation network \rightarrow covariance matrix (1)

- A design choice

(*) GGM - Markov network \rightarrow 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)

- \Rightarrow 2 variables independent given the other variables

1) correlation network:-

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



2) C.I / pcc offer better dep measure

$$Q_{ij,j}=0 \Rightarrow X_i \perp X_j | X_{-ij} \quad \text{or} \quad p(X_i, X_j | X_{-ij}) = p(X_i | X_{-ij})p(X_j | X_{-ij})$$

(*) Sparsity

- later

(*) Network structure learning with LASSO

②: overall
Argument
here

- ex: is there a false distinction between CN and MN?

- after all, precision matrix and covariance matrix are mathematically related via inversion

↓
↳ depends on matrix properties governing possibility of inversion
(e.g. singularity, positive definiteness, rank, eigenvalues) ✓

(*) Assuming covariance invertible, introduces hidden assumptions

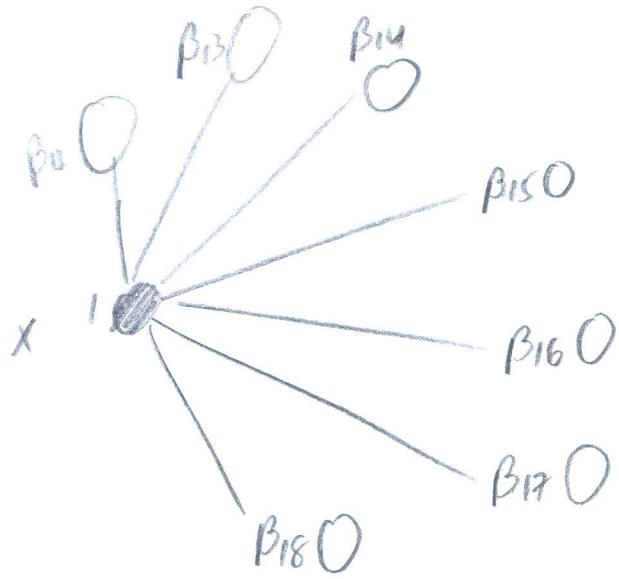
- full rank covariance matrix

- rank deficiency \Rightarrow non invertible covariance $\not\rightarrow$ negation of MV as a model.

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

↳ estimation of Q can be done piece-wise algorithmically.

(*) Piecewise algorithmic, learn structure incrementally.
(try this)



$$x_i = \beta x_i + \epsilon$$

- LASSO regression of all nodes to a target node.

- assume network is a GLM.

- 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: recompile questions

↳ does algorithm yield a graph?

- CS/engineering approach
(more break-ad-build)

↳ is the graph 'correct'?

(*) operational problem in above thought experiment.

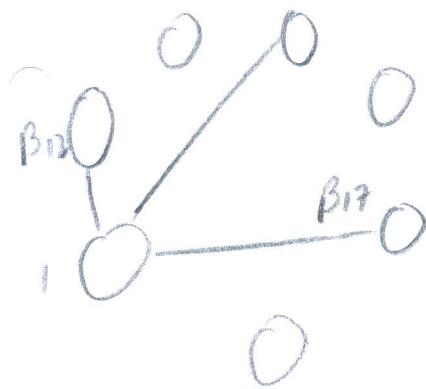
- regression will not yield non-zero β_i

- all nodes will be your neighbor \rightarrow fully connected graph

- so how to avoid to get a meaningful graph

(*) Add sparse regularization (ℓ_1)

- $\hat{\beta}_1 = \arg \min_{\beta_1} \|Y - X\beta_1\|_2^2 + \lambda \|\beta_1\|_1$



(*) ℓ_1 regularisation

- sparsity enforcing constraint via ℓ_1 norm (diamond)

ℓ_2 norm (circle) (no sparsity)

ℓ_p ...

(*) Recall why ℓ_1 norm \rightarrow sparsity

(*) ℓ_0 norm - spikes \rightarrow non-convex; difficultly using gradients.

- find/solve using Lagrangians

- also known as neighborhood selection

(*) There will be some edges non-zero; rest pushed to zero (sparsity)

(*) network learning with LASSO

✓ introduce a rule to encode intuition @1

- 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. ✓
- (*) Now we introduce above rule
 - edge features if one of the regression coefficient estimates is
i.e. neither edge
 $\underline{\text{not zero}}$ *partial slope* \Rightarrow C.I.
- ex: whether this a MRF consistent with previous definition \rightarrow
a concern about validity of graph based on definition
- ex: neighbourhood selection is extremely simple computational-complexity
nB3C.

consistent structure Recovery (2006)

- (*) consistent structure Recovery (2006, Meinshausen & Bühlmann)
- ex: A evolutionary finding (Wainwright 2006, Meinshausen & Bühlmann)
 - under certain conditions; graph estimated by neighbourhood selection /
 LASSO is the same graph described as GLM works to precision
 matrix of GLM.
 - for continuous r.v.
- ①: If $\lambda_S > C \sqrt{\frac{\log p}{S}}$ then with high prob.
 $s(\hat{\beta}) \rightarrow s(\beta^*)$

- (*) Read papers structure disc.
- (*) provide intuition on why algorithm works
- (*) repeated application of neighbourhood selection yields
i.e. LASSO reg
 paper graphical model consistent with GLM.

MVH:

- some interesting properties.

- use some partitioning
- ⑥: remember Bishop/Murphy \rightarrow conditional, marginal Gaussians + matrix inversion lemma.
- $P\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} | M, \Sigma\right) = N\left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} | \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$

- joint Gaussian over substructures induced by part.

- ⑦: How to write down $p(z_2)$, $p(z_1|z_2)$, $p(z_2|z_1)$?

- which satisfies theoretical properties

$$p(z_2) = N(z_2 | M_2^m, V_2^m) \quad p(z_1|z_2) = N(z_1 | M_{12}, V_{12})$$

$$M_2^m = \mu_2$$

$$V_2^m = \Sigma_{22}$$

$$M_{12} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (z_2 - \mu_2)$$

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

adjusted μ_1 and Σ_{11}
with offset terms.

- used a lot when dealing with linear systems.

ex: commit to memory

(4) matrix-inversion lemma

⑦ - Review

(5) 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 distri of one node given all other nodes IF the whole distri is Gaussian

(6) $p(x_i | X_{-i})$

- w/o loss let $\mu = 0$

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

$$(*) p(x_i | \bar{x}_i) = N\left(\frac{\bar{q}_i^T}{q_{ii}} \bar{x}_i, q_{ii}^{-1}\right)$$

- Note the q terms?

(*) Ex: If you see conditional distri with Gaussian mean ad covariance?

conditional auto-reg.

$$x_i = \theta \bar{x}_i + \epsilon \quad \epsilon \sim N(0, \sigma^2) \Rightarrow x_i = \frac{\bar{q}_i^T}{q_{ii}} \bar{x}_i + \epsilon$$

(*) neighborhood selection \rightarrow estimate θ \rightarrow use rule to pose whether non-zero (i.e. care only about structure)

$$- p(x_i | \bar{x}_i) = N\left(\frac{\bar{q}_i^T}{q_{ii}} \bar{x}_i, q_{ii}^{-1}\right)$$

encodes MN
of GGM

$$Q = \begin{bmatrix} & & & \\ & \ddots & & \\ & & \boxed{\text{I}} & \\ & & & \ddots \\ & & & & \end{bmatrix} \quad \frac{\bar{q}_i^T}{q_{ii}} = \begin{bmatrix} & & \\ & & \\ & & \boxed{1} \\ & & \\ & & \end{bmatrix} \quad \hat{\theta} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \hat{\theta}_1 \\ 0 \\ 0 \end{bmatrix}$$

(*) sparseness of θ is the same as sparseness of Q ✓ (i.e. forget values only focus on the component in θ is zero ($\Rightarrow Q$ comp is zero))
(one-to-one correspondence?)

Ex: If our goal is to estimate sparseness of θ and Q ; i.e. the parameter locations that encode structure; then this achieves goal.

$$S_i = \{j : j \neq i, \theta_{ij} \neq 0\} \quad \text{④ : ②} \quad \checkmark$$

Ex: The maths is not the key message; rather insights / angles of viewing.
(*) - Do it for all $x_i, i=1, \dots, N$
- to get Q (also symmetric)

- story

(*) Each estimation $\frac{\bar{q}_i^T}{q_{ii}}$ (via regression of x_i on \bar{x}_i) yields an LASSO. estimate of a column of precision matrix

(*) Meinschusen-Buhmann algorithm

(*) L₁-regularised ML learning

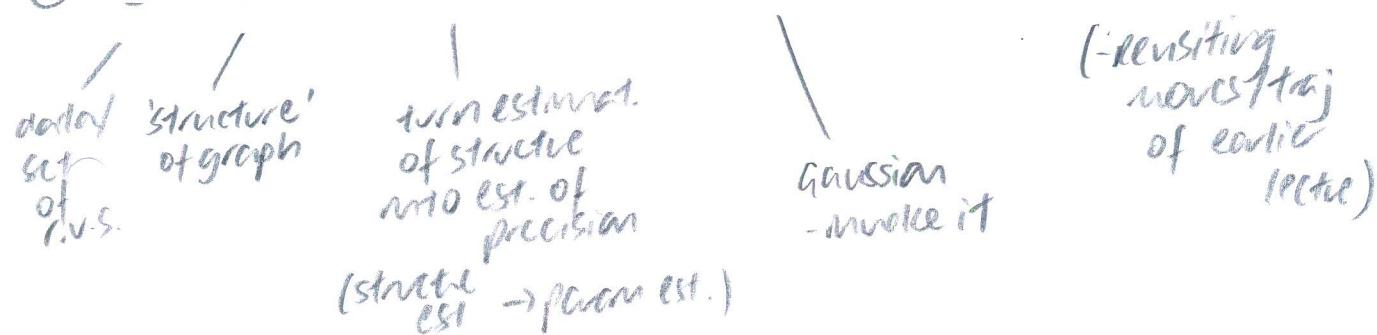
(*) ...

(*) learning Ising models (pairwise MRF)

- assuming discrete valued nodes

• can we turn a structure learning problem into a param estim. problem?

• $\mathcal{X} \rightarrow S_\lambda \rightarrow S \equiv Q \rightarrow \Sigma^{-1} \rightarrow N(\mu, \Sigma^{-1}) \rightarrow p(x_i | x_{-i}) \rightarrow \text{regression}$



- some of these moves heuristic, some mathematical.

$$p(\Sigma | \Theta) = \exp \left(\sum_{i \in V} \theta_{ii}^T x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j - A(\Theta) \right)$$

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

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

Q: → empirical distri (i.e. counting)?

Ex: Hint: invoke linear regression for continuous case as a placeholder for problem → after placeholder

- modify to do a regression on discrete-rvs. → logistic regression

(*) Build this kind of insight (heuristic)

$$p(x_K | x_{-K}) = \text{logistic} \left(2x_K \langle \theta_{1K}, x_{-K} \rangle \right) \text{ with } L_1\text{-regularisation.}$$

(*) Proof → KKT / complementarity etc.

(*) scalar \rightarrow vector valued.

- ex: what is hidden assumption when moving from scalar to vector
I.V.S.?

- IID validity - IID over scalars vs vectors

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

- modify to allow vector valued output. (Bishop)

- linear/logistic \rightarrow scale output

(*) vector valued nodes

- ① How to deal with vectors?

- ② How to deal with vectors of variable length? \rightarrow kernels on structured data ②

ex: possible to directly estimate partial correlations of vectors of different lengths etc.

Q: How to estimate evolving networks?

- estimate a network from one bucket of votes at the t?

- statistical estimates need to be built on redundancy of data?

- You have only one data point

- IID is violated

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

(*) inferentia 1

(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 lin. 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
 \cap an explicit

- difficult to NN

(*) Inference II - skipped

- need literature on how heuristic + theory work together.

(*) Rest of slides skipped.

- state of art on learning network structure

(*) consistent, mathematically provable structure.