Graphical Modelling Guided Study

Updated: 7.4.2023

Preface

- Exercises are presented with omissions, solutions are unverified
- Proofs are similarly good-faith efforts but unverified
- Red text indicates my personal questions, lapses in understanding, or otherwise shakey areas
- Other nice treatments of similar material include:

  Frederic Koehler's lecture on Common Gaussian Graphical Models https://www.youtube.com/watch?v=V6NMDZB6LI4

  – Illinois lecture note on graphical models class: http://swoh.web.engr.illinois.edu/courses/IE598/info.html
- Useful[1] references/notes on convex optimization and sub-gradient notation

  Ryan Tibshirani's Convex Optimization Notes at https://www.stat.cmu.edu/~ryantibs/convexopt/

  Boyd and Vandenberghe's 2008 Textbook on Convex Optimization https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf

[1] Not a pre-req (or any area of experience for me) but relevant optimization tools do crop up in Ch. 17 of ESL

To Do

Important

- Summarize/scribe notes on Chen et al paper
- Finish ESL Ch. 17 Review
- Implement Graphical Lasso (ESL Exercise 17.8)

Supplementary

- Review Wasserman Ch 19 (log-linear models)
- Review proofs:

  Hammersley-Clifford theorem (iff/bi-directional)

  Markov Property: Global $\Leftrightarrow$ Pairwise
- Review Junction-Tree Algo (ESL Ch 17)
- Review Wasserman Ch 19 (log-linear models)
- Treatments of Graphical Grouped Lasso
  - https://www.jmlr.org/papers/volume21/19-181/19-181.pdf
  - https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2015/2010-Friedman-Hastie-Tibshirani.pdf
- Undirected (Graphical) Elastic Net https://arxiv.org/abs/1111.0559

Misc. Proofs

Publications

Chen (2014): *Selection and Estimation for Mixed Graphical Models*

Takeaways/High-Level Notes

- This paper extends previous work to allow for estimation of conditional dependencies/associations within graphs of mixed distributions within the exponential family. Previous work focused on Gaussian graphs and more recently within

  One related/contemporaneous work allowed for a mixed model of two distributions, this work allows for any(?) combination of the specified exponential family distributions

Further Review

- Work through derivations of Ex. 1 to 4 (parameterization of conditional densities $p(x_s|x_{-s})$ in the proposed conditional density form (3))
  - See http://www.cs.cmu.edu/~epxing/Class/10708-16/note/10708_scribe_lecture10.pdf
- Revisit equation (3) (pg 3), how was this derived or arrived at? Is this a known extension of the exponential family definition in the graphical setting?
- Review proofs in Appendix

Notes/Questions

- The Introduction mentioned papers from ∼2009-13 that proposed semi-/non-parametric methods for conditional dependence estimation (Graphical Random Forest; Joint Additive Models) but criticizes these methods' efficiency. Is non-parametric estimation still an open research area?
- What are these node potential functions, $f(x_s)$? Are they present to capture/account for the marginal "densities"(?) of a given random vector/node x_s? I want to understand because they seem to define the importance of α_s, which in turn are parameters that we assume are known in the algorithm proposed in Section 3. Just trying to understand 1) what we are estimating in estimating α_{s1} and how strict (or just what the) assumption is when we say α_{sk} are known for $k \geq 2$

  The α_s are defined as $f(x_s) = \alpha_{s1}x_s + \alpha_{s2}x_s^2/2 + \sum \alpha_{sk}B_{sk}(x_k)$. That is, the α_s vector are coefficients for some linear combination that defines the node potential function
- Why allow for different penalty λ by node type? This was counterintuitive to me. We've assumed that our graph is undirected, or $\theta_{st} = \theta)_{ts}$. However if x_s, x_t are distributed differently (e.g. one Poisson and one binomial), $\not\Rightarrow \lambda_s\theta_{st} = \lambda_t\theta_{ts}$ (we could apply different optimal λ values

- General question: Neighbourhood selection (and Graphical Lasso) are defined on ℓ_1 penalty alone. Are there extensions (and if so, are they popular/open reserach) on ℓ_2 or combined penalties?

  Found a 2011 treatment of an elastic net model for undirected Gaussian undirected graphs out of Princeton at https://arxiv.org/abs/1111.0559

- Very general question: How (If you do) do you pick out useful techniques from some of the dense theory? And how do you identify what is a necessary pre-requisite or where you can assume or glance over some knowledge? For example, I don't have a convex optimization background outside of it's discussion in some stat theory (read: very little). So the proof of theorem 1 relying primal-dual witness method. Is this worthwhile to learn and how do I identify what is a useful/generalizable tool/skill and what may be useful just for this paper (and worth either accepting, or asking someone else to verify that I do know has this background)?

Wainwright (2012) *Unified Framework for High-Dimensional Analyais of M-Estimators with Decomposable Regulators*

Further Review/Questions

- Where does equation 18 come from? I am familiar with explicit forms of the remainder in univariate expansions, but I'm not familiar with the multivariate remainder expression (and am unsure how this specific expression is derived)
- Review beginning of 4.2, 1) do I understand this Taylor series expansion about $\partial\mathcal{L}$? 2) do I understand why it is exact in this scenario?
- End of page 12, how does this correspond to a restricted eigenvalue condition?

Takeaways

- 4.1) For least squares LASSO (ℓ_1 penalty), the Taylor series involved in the RSC definition is exact, and thus independent of θ^* (the parameter's true value for all intents and purpose). This, combined with the definition of the cone set $\mathcal{C}$ to which $\hat{\Delta}$ must belong simplifies the necessary RSC demonstration into a restricted eigenvalue condition, which can be shown to be met with high probability for Gaussian (and subGaussian) design matrices (even with dependencies)

  But no comments yet on convergence, accuracy, etc. just meeting this RSC definition as necessary for further analysis of this M-estimation problem

- 4.2) Assume RE to assume RSC. We know ℓ_1 is decomposable. Thus we have the bounds from Corollary 1 on the error vector, we must simply identify the regularization constant λ_n and the compatibility function $\Psi(M)$ for $\mathcal{R}(\cdot) = ||\cdot||_1$ and error norm $||\cdot||_2$ (which is the meat of the proof of Corollary 2, besides the implicit $RE \Rightarrow RSC$ argument that I believe 4.1 makes)

  Note the sub-gaussian (and normalizing) assumption allows for the high-probability argument made in the proof of corollary 2.

- 4.3) So assuming now θ^* is weakly sparse, that is not sparse but adequately-approximated by a sparse vector. We construct a set of sparse vectors for $\mathcal{M}$, and now have that $\theta^* \notin \mathcal{M}$, leading to 1) this ball-shaped set and 2) the need for a positive tolerance function $\tau(\theta^*)$. This is contradicted however by the statement in Corollary 3 that $\theta^* \in \mathbb{B}_q(R_q)$, that is θ^* actually does belong to our model/sparsifiable set.

  Maybe it's that θ^* belongs to $\mathbb{B}_q(R_q)$ but we are estimating with a truly sparse set (i.e. $\mathbb{B}_0(R_0)$ with at most R_0 non-zero features), a stricter condition/subset of the $\mathbb{B}_q$ sparsifiable set

Definitions in section 2 as set-up for Theorem 1:

1. Identifies a suitable bound on our penalty λ_n related to the norm $\mathcal{R}(\nabla\mathcal{L}(\theta^*))$ such that our errors δ for any optimal solution $\hat{\theta}$ are contained in a bounded cone (if $\theta^* \in \mathcal{M}$, a star-shaped set otherwise)
2. Strong convexity of our loss function $\mathcal{L}$ is necessary to ensure that $|\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta}_{\lambda_n}|$ approaching 0 implies that δ (or $\hat{\delta}$ is also small. We can be less strict in this assumption and only require strong convexity in a neighborhood about θ^*, and in fact can focus on the cone (or star-shaped set) identified in definition 1, leading us to this RSC definition/consideration

3. It seems intuitive to want some subspace compatibility measure, as we can construct any subspace M (and norm, with decomposability as in Dfn 1) and $\overline{M}^{\perp}$ that we deem appropriate, but what is the idea behind this specific measure of compatibility? My intuition is that this captures the interplay between the subspace M (by suping over $u \in M$) and balancing the norm $\mathcal{R}$ with the

- namely, a decomposability property for the regularizer and a notion of restricted strong convexity that depends on the inter- action between the regularizer and the loss function.

- Decomposability (want to ensure I understand this perturbation intuition):

  Consider a model M and our estimation $\overline{M}$ (for simplicity consider $M = \overline{M}$ but general result offered for $\overline{M} \subseteq M$. For $\theta \in M$ and $\gamma \in \overline{M}$, $\theta + \gamma$ (and specifically γ) can be considered deviations away from the model space

Questions/Notes

- How do you best understand/learn the "history" of a given topic/method?

  Sometimes it frequently pops up and just sticks (Tibshirani 1996 key ex), either in coursework or a common "parent" publication across

  "Review"-esque papers like this (or at least papers that provide succinct overviews of notable papers) are great but not always available

- How do we think about $\theta^* \in M$? Is this somewhat similar to identifiability? Like a well-specified model, such that M our subspace which encodes the constraints (or other parameterizations/properties) that we believe are true for θ actually contains θ^* (analogous to "ground truth" in the frequentist univariate parameter estimation setting)

- **More transparently, what is $\overline{M}$ in relationship to M?** Topological closures(?): Important to understand the distinction $\overline{M} \subseteq M$ and $\overline{M} = M$? I understand when we do not have equality (low-rank matrix estimation), but I don't necessarily understand what a topological closure is

- (pg. 5-6) *"With certain exceptions, it is computationally expensive to enforce a rank-constraint in a direct manner..."* What are the issues? Something related to the number combinations of row/column combinations that could result in a specific rank become prohibitively large (even if considered cleverly)?

- (pg. 11 (c)) How does the tolerance $\tau_{\mathcal{L}}$ related to unidentifiable components in a high-dim model?

Further Review

Pressing/Paper Material

- Figure 1 (3-dimensional error vector, geometric intuition behind $\mathbb{C}(M, \overline{M}^{\perp}, \theta)$ when $\theta^* \in M, \theta^* \notin M$ respectively

(Newly) Conceptual

- Review equivalency of lasso and basis pursuit de-noising

  See https://www.cs.cornell.edu/courses/cs6220/2017fa/CS6220_Lecture21_2.pdf

  Possibly only of historical relevance

- Definition/abstraction of a (topological) closure:

  For subspace $M \subseteq \mathbb{R}^p$, $\left(M^\perp\right)^\perp \equiv \overline{M}$ is a closure of M (more accurately is a closure operator on M)

  With the exception of discussion of low-rank matrices and the nuclear norm, we can work with $M = \overline{M}$ and not concern ourselves with the concept of topological closures

- Prove (pg 7) dual-norm for group-norm (similar to $\ell_1 - \ell_\infty$ dual norm relationship proof)

Misc (Somewhat) Related Work

Proof of equivalent dual-norm definitions

WTS $\sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u,v \rangle}{\mathcal{R}(u)} = \sup_{\mathcal{R}(u) \leq 1} \langle u, v \rangle$
1)

$$A := \{u : \mathcal{R}(u) \leq 1\} \subseteq \{u : ||u|| \leq 1, u \neq 0\} =: B$$
$$\Rightarrow \sup_{\mathcal{R}(u) \leq 1} \langle u, v \rangle \leq \sup_{\mathcal{R}(u) \leq 1} \frac{\langle u, v \rangle}{\mathcal{R}(u)} \leq \sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{\mathcal{R}(u)}$$

Elements of Statistical Learning

Chapter 17: Undirected Graphical Models

Overview/Intro

1. *Omitting information that is shared with Wasserman chapters, which are more introductory than ESL's discussion of graphical algorithms*
2. ~~We define clique potentials (similar to the Wasserman chapter) as affinities (affine functions?). Since we express the density function as product of clique functions, I assume these must be positive functions defined on each clique. Are there other constraints or definitions about these affinities?~~
   - My (still limited, possibly incorrect) understanding is that these are simply positive functions that are context-specific or user-defined. Wainwright/Jordan describes these as *compatibility functions* which are defined based on a model.
   - An example of a simple compatibility functions is a binary decision rule that is 0 for any configuration of vertices/values that occurs with probability 0, and 1 otherwise. That is for clique (x_i, x_j, x_k) with known impossible configuration (z_i, z_j, z_k), the compatibility function is equivalent to the boolean $\neg z_i \vee \neg z_j \vee \neg z_k$.
3. Hammserley-Clifford: From ESL, states that we can equivalently represent the joint density function of a graph ($\mathcal{G} = (V, E)$) f_V as a product of clique affinities **for Markov networks with positive (i.e. non-zero) distributions**
   - That is for set of maximal cliques[2] $\mathcal{C}$ and graph $\mathcal{G} = (V, E)$, $f_{\mathcal{G}}(x) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$
4. Another question re: HC theorem/clique-factorization of the density function "implies a graph with independence properties defined by the cliques...". So this is true even with overlap in the maximal cliques (or within whatever cliques are used to factorize the density function)?

17.3 Continuous Variables

- Assuming a Gaussian distribution describing our nodes allows for some convenient estimation properties to arise in graph structure and/or parameter estimation:
  - Generally (or perhaps vaguely), Gaussian graphical models allow estimation problems to be constructed conveniently as linear regression problems (see 17.3.1 for regression estimating equations, 17.3.2 for lasso regression for structure estimation)
  - For $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $\theta_{ij} = 0 \Leftrightarrow X_i \perp X_j |$rest
- I assume the algorithm for graphs with completely known structure is a (almost prohibitively) rare setting for any utility. Are there also settings with partially specified structure? Suppose we know the existence/absence of some specific edges but otherwise have no other information? A hybrid of the structure-specified and -unspecified approach?

[2] From Wainwright, Jordan (2008), I believe we can also use a set of non-maximal cliques (or any combination of cliques) in this factorization when convenient

17.3.1 Estimating Equations for Graphs with Known Structure

~~Pg 631, authors claim *"The dependence of Y on Z in (17.6) is in the mean term alone"* for this conditional Gaussian relationship. How is this true? We also see this σ_{XY} term in the Variance as well, where σ_{XY} is the covariance of Y, X.~~ The dependence on Z is only in the mean. The variance term includes σ_{ZY} which is a parameter/"constant" (i.e. does not vary by values of Z, just captures the dependence relationship between Y, Z.

See work/question for Exercise 17.5 for question of distinction between W, S matrices in presented algorithm

17.3.2 Estimation of Graph Structure (Graphical Lasso)

I don't necessarily understand the difference between the initial 2006 lasso approach to structure estimation (fitting p lasso models, arguing edge θ_{ij} exists iff β_j is non-zero for X_i regressed on $\mathbf{X}\backslash\{X_i\}$) and the algorithm described in 17.2 (graphical lasso). GL is solving an equivalent system of equations, no? And this algorithm has guaranteed convergence pending lasso convergence. I would need to read the 2008 graphical lasso paper to undertand its theoretical guarantees and/or possible convergence issues

17.4 Discrete Variables

1. (a) Maximum Cliques: $\{X_1, X_2, X_3\}$, $\{X_1, X_4\}$, $\{X_3, X_4\}$, $\{X_5, X_6\}$

   (b) Conditional Independencies: Trivially, any X_i, X_j without an edge is independent conditional on all other nodes
   By separation, a list (with some redundancies):
   $X_1 \perp X_5 | X_6$
   $X_2 \perp X_{3,4} | X_1$ $X_2 \perp X_6 | X_5$
   $X_3 \perp X_{1,2,5,6} | X_4$
   X_4 & $X_3 \perp X_{2,5,6} | X_1$
   $X_5 \perp X_{1,2,3,4} | X_6$

2. *Omitted*

3. (a) Note $\Sigma \in \mathbb{R}^{p \times p}$ can be partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

   for $\Sigma_{aa} \in \mathbb{R}^{2\times 2}$; $\Sigma_{ab}, \Sigma_{ba}^T \in \mathbb{R}^{2 \times p-2}$; $\Sigma_{bb} \in \mathbb{R}^{p-2 \times p-2}$.
   We can partition $\Theta \equiv \Sigma^{-1}$ and use known properties of the inverses of partitioned matrices to demonstrate:

$$\Theta = \begin{bmatrix} \Theta_{aa} & \Theta_{ab} \\ \Theta_{ba} & \Theta_{bb} \end{bmatrix} = \Sigma^{-1} = \begin{bmatrix} \left(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1} & f_2(\Sigma) \\ f_3(\Sigma) & f_4(\Sigma) \end{bmatrix}$$

   where we see $\Theta_{aa} = \Sigma_{a,b}^{-1}$.

   (b)

$$\Sigma_{a,b} = \begin{bmatrix} Cov(X_1, X_1|\text{rest}) & Cov(X_1, X_2|\text{rest}) \\ Cov(X_2, X_1|\text{rest}) & Cov(X_2, X_2|\text{rest}) \end{bmatrix}$$
$$\Rightarrow \Sigma_{a,b}^{-1} = \Theta_{aa} \propto \begin{bmatrix} Cov(X_2, X_2|\text{rest}) & -Cov(X_1, X_2|\text{rest}) \\ -Cov(X_2, X_1|\text{rest}) & Cov(X_1, X_1|\text{rest}) \end{bmatrix}$$

   The off-diagonals of $\Theta_{aa} = 0 \Rightarrow \rho_{1,2|\text{rest}} = 0$. Noting that we've selected a, b such that $X_a = (X_1, X_2)$ as $j = 1, k = 2$ WLOG (i.e. the result holds $\forall j \neq k$) completes the argument.

   (c) [3] We can (less lazily compared to 3b) calculate the partition of precision matrix Θ_{aa}, where $\theta_{ij} = Cov(X_i, X_j|\text{rest})$:

$$\Sigma_{a,b}^{-1} = \Theta_{a,b} = \frac{1}{\theta_{ii}\theta_{jj} - \theta_{ij}^2} \begin{bmatrix} \theta_{jj} & -\theta_{ij} \\ -\theta_{ji} & \theta_{ii} \end{bmatrix}$$

   Then $\text{diag}(\Theta)^{1/2} = \begin{bmatrix} \frac{1}{\sqrt{\theta_{jj}}} & 0 \\ 0 & \frac{1}{\sqrt{\theta_{ii}}} \end{bmatrix}$ and:

[3] Wikipedia actually has a nice walkthrough of the calculation of the partial conditional correlation formula, at https://en.wikipedia.org/wiki/Partial_correlation#Using_matrix_inversion

$$\mathbf{R} = \text{diag}(\Theta)^{-1/2} \cdot \Theta \cdot \text{diag}(\Theta)^{-1/2} = \begin{bmatrix} \sqrt{\theta_{jj}} & -\frac{\theta_{ij}}{\sqrt{\theta_{jj}}} \\ -\frac{\theta_{ji}}{\sqrt{\theta_{ii}}} & \sqrt{\theta_{ii}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\theta_{jj}}} & 0 \\ 0 & \frac{1}{\sqrt{\theta_{ii}}} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & -\frac{\theta_{ij}}{\sqrt{\theta_{jj}\theta_{ii}}} \\ -\frac{\theta_{ji}}{\sqrt{\theta_{jj}\theta_{ii}}} & 1 \end{bmatrix}$$

where we see $r_{ij} = \rho_{ij}|\text{rest}$ by definition of $\rho_{ij}|\text{rest} = -\frac{\theta_{ij}}{\sqrt{\theta_{jj}\theta_{ii}}}$.

4. Notation: Let $\{X_3, X_4, ..., X_p\} = X^* = X_{3,...,p}$:

$$X_1 \perp X_2|X^* \Leftrightarrow f_{X_1,X_2|X^*} = f_{X_1|X^*} f_{X_2|X^*}$$
$$f_{X_1,X_2|X^*} = \frac{f_{X_1,X_2,X^*}}{f_{X^*}} = \frac{f_{X_1|X_2,X^*} f_{X_2|X^*} f_{X^*}}{f_{X^*}}$$
$$= f_{X_1|X_2,X^*} f_{X_2|X^*}$$
$$= f_{X_1|X^*} f_{X_2|X^*}$$

5. From 17.3.1 (and work below under Misc. Claims), the gradient of the log-likelihood for our Gaussian graphical model is $\nabla \ell(\Theta; \mathbf{X}) = \Theta^{-1} - S$ (with $\Gamma = \mathbf{0}$, as all edges are known and present). $S = \Theta^{-1}$ and assuming a similar partitioning scheme as in 17.3.1:

$$S = \Theta^{-1} \Rightarrow \begin{bmatrix} S_{11} & s_{12} \\ s_{21} & s_{11} \end{bmatrix} \begin{bmatrix} \Theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & 0 \\ 0^T & 1 \end{bmatrix}$$
$$\Rightarrow S_{11}\theta_{12} + s_{12}\theta_{22} = 0$$
$$\Rightarrow s_{12} = -\frac{S_{11}\theta_{12}}{\theta_{22}} = S_{11}\beta$$

$$\Theta - S \stackrel{!}{=} 0 \Rightarrow s_{11} - s_{12} = 0 \Leftrightarrow S_{12}\beta - s_{12} = 0 \quad \square$$

This problem feels a bit weird. In order to use this substitution, the gradient gives us $s_{12} - s_{12}$, no? Which is trivially true. IN the 17.3.1 derivation, we use W. Not sure if I truly understand the distinction of S, W.

6. Omitted, result follow nearly immediately from 17.16 (or the provided 17.41-2) and a profile-likelihood style argument

7. *Incomplete (Programming)*

8. *Incomplete (Programming)*

9.

Misc. Claims, Proofs, Work

Claim: The log-likelihood of N random samples of a $k-$dimenisonal multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}$ can be expressed as (noting $\mathbf{x} \in \mathbb{R}^k; \boldsymbol{\Theta}, \mathbf{S} \in \mathbb{R}^{k \times k}$) for sample covariance matrix $\mathbf{S} = (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T$:

$$\text{WTS} \ell(\boldsymbol{\Theta}) = (\propto ?) \log\det \boldsymbol{\Theta} - \text{trace}(\mathbf{S}\boldsymbol{\Theta})$$

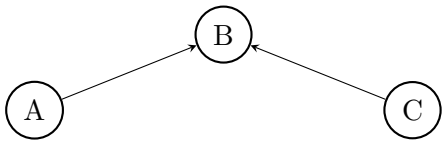
$$\begin{aligned}
\ell(\boldsymbol{\Theta}) &= \sum_{i=1}^{N} \log \left[(2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-k/2} \det(\boldsymbol{\Theta})^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Theta} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right] \\
&\propto \log\det \boldsymbol{\Theta} - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Theta} (\mathbf{x}_i - \boldsymbol{\mu}) \\
&\overset{\overline{\mathbf{x}} = \hat{\mu}_{MLE}}{\propto} \log\det \boldsymbol{\Theta} - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})^T \boldsymbol{\Theta} (\mathbf{x}_i - \overline{\mathbf{x}}) \\
&\overset{?}{=} \log\det \boldsymbol{\Theta} - \text{trace}(\mathbf{S}\boldsymbol{\Theta})
\end{aligned}$$

All of Statistics (Wasserman)

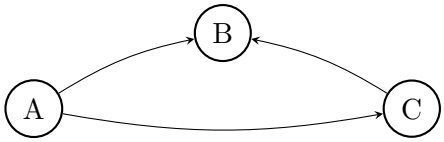
Chapter 17: Directed Graphs

Questions, Definitions, Notes, Properties, etc.

1. An **unshielded collider** is any collider whose "pointing nodes" are disconnected/non-adjacent:



is unshielded, whereas



is shielded

2. A distribution $\mathbb{P}$ for nodes $V = \{X_1, ..., X_k\}$ is Markov wrt a graph $\mathcal{G}$ if $f(v) = \prod_{i=1}^{k} f(x_i|\pi_i)$ for π_i parents for node X_i. Also written as $\mathbb{P} \in M(\mathcal{G})$
3. The **Markov Condition** for distribution $\mathbb{P}$ holds if $\forall X_i \in V, \mathcal{G} = (V, E)$ (or for X_i simply as random variables) if $W \perp \tilde{W}|\pi_W$, where $\tilde{W}$ includes all other nodes/variables besides π_W and descendants of W
4. The following items from this list are equivalent characterizations $\mathcal{G}$: 2 $\Leftrightarrow$ 3
5. For disjoint sets of verticies A, B, C: A, B are d-separated by $C \Leftrightarrow A \perp B|C$
6. $\mathcal{G}_1, \mathcal{G}_2$ are **Markov Equivalent** $\Leftrightarrow \mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2) \Leftrightarrow$ skeleton$(\mathcal{G}_1)$ = skeleton$(\mathcal{G}_2) \wedge$ both graphs have the same unshielded colliders

Exercises

***** 1

WTS (17.1) and (17.2) are equivalent: $X \perp Y|Z$ indicates $f_{X,Y|Z} = f_{X|Z} f_{Y|Z} \Leftrightarrow f_{X|Y,Z} = f_{X|Z}$

$$f_{X|Y,Z} = \frac{f_{X,Y,Z}}{f_{Y,Z}} = \frac{f_{X,Y,Z}}{f_{Y|Z} f_Z} = \frac{f_{X,Y|Z}}{f_{Y|Z}} \overset{X \perp Y|Z}{=} \frac{f_{X|Z} f_{Y|Z}}{f_{Y|Z}} = f_{X|Z}$$

***** 2

$$\mathbb{P}(U \leq u, Y \leq y|Z) = \mathbb{P}(X \leq h^{-1}(u), Y \leq y|Z) = \mathbb{P}(X \leq h^{-1}(u)|Z)\mathbb{P}(Y \leq y|Z) = \mathbb{P}(U \leq u|Z)\mathbb{P}(Y \leq y|Z)$$

(a) *Trivial*

(b) WTS $X \perp Y|Z \wedge U = h(X) \Rightarrow U \perp Y|Z$:

$$f_{U,Y|Z}(u, y) = f_{X,Y|Z}(h^{-1}(u), y)\left|\frac{\partial h^{-1}(u)}{\partial u}\right| = f_{X|Z}(h^{-1}(u))\left|\frac{\partial h^{-1}(u)}{\partial u}\right| f_{Y|Z}(y) = f_{U|Z}(u)f_{Y|Z}(y)$$

(c) WTS $X \perp Y|Z \wedge\ U = h(X) \Rightarrow X \perp Y|(Z, U)$:

$$f_{X,Y|Z,U} = f_{Y|X,U,Z}f_{X|U,Z} \overset{U=h(X)}{=} f_{Y|U,Z}f_{X|U,Z}$$

(d) WTS $X \perp Y|Z \wedge X \perp W|(Y, Z) \Rightarrow X \perp (W, Y)|Z$

$$f_{X,W,Y|Z} = f_{W|X,Y,Z}f_{X,Y|Z} \overset{X\perp W|(Y,Z)}{=} f_{W|Y,Z}f_{X|Y,Z} \overset{X\perp Y|Z}{=} f_{W|Y,Z}f_{X|Z}$$

(e) WTS $X \perp Y|Z \wedge X \perp Z|Y \Rightarrow X \perp (Y, Z)$ (without assumption of positivity for all involved probabilities)

$$f_{X,Y,Z} = f_{Z,Y|X}f_X = f$$

***** 3

Omitted

***** 4

Consider the (re-created) DAG's in 17.6 with no colliders present:

$$\mathrm{X} \longrightarrow \mathrm{Y} \longrightarrow \mathrm{Z}$$

$$\mathrm{X} \longleftarrow \mathrm{Y} \longleftarrow \mathrm{Z}$$

$$\mathrm{X} \longleftarrow \mathrm{Y} \longrightarrow \mathrm{Z}$$

WTS $X \perp Z|Y$

1. The Markov Condition directly implies $Z \perp X|Y$ (Z is independent of all nodes excluding its parents $\{Y\}$ and descendants $\{\varnothing\}$ conditioned upon its parents)
2. $X \perp Z|Y$ again by Markov condition (same as above)

3. Markov Condition again in a similar way wrt either X, Z (both have empty set descendants, Y as parent)

* 5

$$\mathrm{X} \longrightarrow \mathrm{Y} \longleftarrow \mathrm{Z}$$

Consider now the above DAG with a collider present, WTS $X \perp Z$ and $X \not\perp Z|Y$:

$X \perp Z$ follows from Markov Condition (Z is not a descendant of X, X has no parental nodes) or noting that X, Z are d-separated (specifically only when **not** conditioning on Y).

We know that $X \perp Z|Y \Leftrightarrow X, Z$ are d-separated. We note by definition X, Z are d-connected conditioning on Y and thus $X \not\perp Z|Y$.

* 6

*S*imulations omitted

$$f_{X,Y,Z} = f_{Z|Y} f_{Y|X} f_X$$

* 7

DAG Omitted
Consider the set of nodes $V = \{Z_j, X, Y_i\}$, $i, j = [4]$:

$$f_V = f_X \prod_{k=1}^{4} f_{Z_k} f_{Y_k|Z_k,X}$$

$X \perp Z_j, \forall j \in [4]$ follows directly from the Markov Condition, as no Z_j is a parent or descendent of X. We could also note X, Z_j collide at Y_j and are d-separated, thus $X \perp Z_j$ but $X \not\perp Z_j|Y_j$ ($\forall j \in [4]$).

* 8

(a)

$$\mathbb{P}(Z|Y) = \frac{\sum_{x=0}^{1} \mathbb{P}(Z, Y, X = x)}{\sum_{x=0}^{1} \mathbb{P}(Y, X = x)} = \frac{\sum_{x=0}^{1} \mathbb{P}(Z|Y, X = x)\mathbb{P}(Y|X = x)\mathbb{P}(X = x)}{\sum_{x=0}^{1} \mathbb{P}(Y, X = x)}$$

Result omitted, calculation follows from expression above (all information known from given information)

(b) Omitted

(c) *Incomplete*

(d) Omitted

* 9

(a) *Incomplete*

[

Chapter 18: Undirected Graphs]

Questions, Definitions, Notes, Properties, etc.

1. Pairwise Markov property for $\mathcal{G} = (V, E)$, $X, Y \subseteq V$, and $V \backslash \{X, Y\}$ is all nodes excluding X, Y:
$$\text{No edge exists between } X, Y \Leftrightarrow X \perp Y | V \backslash \{X, Y\}$$
2. The Global Markov states for sets of vertices $A, B, C \subseteq V$ in graph $\mathcal{G}$:
$$A \perp B | C \Leftrightarrow C \text{ separates } A, B$$
3. $M_{\text{pair}}(\mathcal{G}) = M_{\text{global}}(\mathcal{G})$

Exercises

* 1

A) $X_1 - X_2 - X_3$ B) $X_1 \quad X_2 - X_3$ C) $X_1 \quad X_2 \quad X_3$

All three relationships also hold trivially for the graph in (C).

* 2

A) $X_1 - X_2 - X_3 - X_4$

B)
$$\begin{array}{ccccc} X_1 & - & X_4 & - & X_2 \\ & & | & & \\ & & X_3 & & \end{array}$$

C) $X_2 - X_3 - X_4 - X_1$

* 3

(a) $X_1 \perp \{X_3, X_4\} | X_2$;
$X_3 \perp X_4 | X_2$

(b) $X_1 \perp \{X_3, X_4\}|X_2$ or $X_1 \perp X_4|X_3$;
$X_2 \perp X_4|X_3$

(c) $X_1 \perp X_3|X_2, X_4$;
$X_2 \perp X_4|X_1, X_3$

(d) $X_1 \perp \{X_4, X_5, X_6\}|X_2, X_3$;
$X_2 \perp X_6|X_3, X_5$; $X_3 \perp X_4|X_2, X_5$;
$X_4 \perp \{X_3, X_6\}|X_2, X_5$;
$X_6 \perp \{X_2, X_5\}|X_3, X_5$

***** 4 *Omitted*

***** 5 *Omitted*