Graphical Modelling Guided Study

Dominic DiSanto working with Junwei Lu

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

Non-Urgent/Of-Interest Review

- Community-Based Group Graphical Lasso (Pircalabelu, 2020) https://www.jmlr.org/papers/volume21/19-181/19-181.pdf
- Applications of lasso/grouped lasso (Friedman 2010) https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2015/2010-Friedman-Hastie-Tibshirani.pdf
- Elastic net model selection in undirected graphical models (Cucuringu 2011) https://arxiv.org/abs/1111.0559
- Review Wasserman Ch 19 (log-linear models)
- Review Junction-Tree Algo (ESL Ch 17 references)
- Original group lasso paper (Yuan, 2006) http://www.columbia.edu/~my2550/papers/glasso.final.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

Notation

Abbreviated, contains only notation that is ambiguous or otherwise necessitates explicit definition

- $0 \in \mathbb{N}, 0 \notin \mathbb{N}^+$
- We abbreviate integer sets as $[k] = \{1, 2, ..., k\}$ (implicitly assertes $k \in \mathbb{N}^+$)
- For any $f : \mathbb{R}^n \mapsto \mathbb{R}$, $\nabla f, \nabla^2 f$ represent the gradient and Hessian respectively. $H_f, \mathcal{H}_f$ may similarly represent the Hessian but I try to minimize use of this notation in favor of $\nabla^2 f$

To Do

Important

- Review Lemma 2 proof
- Review proposition 2 proof (v1 doc)
- Understand theorem(s)/cited papers used to determine that Gaussian/subGaussian design matrices in linear setting

  Exact sparsity, Gaussian-ensemble matrices follows directly from Raskutti 2010

  Exact sparsity, sub-Gaussian Seems like it follows from sub-Gaussian ensemble satisfying the UUP via Mendelson 2006, Uniform uncertainty principle for Bernoulli and subgaussian ensembles but may also need Zhou 2009, Restricted Eigenvalue Conditions on Subgaussian Random Matrices (although this may just be an extension beyond independence)

  Weak sparsity, Gaussian ensemble

  Weak sparsity, sub-Gaussian ensemble

- Find/Prove sub-Gaussian definition (pg. 13):
  $w \sim subG(\sigma^2)$ if $\mathbb{P}(|\langle w, v\rangle| \geq t) \leq 2 \exp\left\{-\frac{t^2}{2\sigma^2}\right\}, \forall ||v||_2 \leq 1$
-

Supplementary

- Implement Graphical Lasso (ESL Exercise 17.8)
- TeX up/finish ESL Ch. 17 notes
- Review proofs:
  - Hammersley-Clifford theorem
  - Markov Properties
    - Global $\Leftrightarrow$ Pairwise $\Leftrightarrow$ Local (for positive distributions)
    - Global $\Rightarrow$ Local $\Rightarrow$ Pairwise (generally)

Misc. Proofs

Hammersley-Clifford Theorem

Equivalence of Pairwise and Global Markov Factorizations of Graph

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?

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
  - To this coefficient point, most often α_{s1}/linear term coefficient is the only (or most important) parameter of interest. This expression of $f(x_s)$ allows for generalizations to include higher order terms, but as the authors note, $\alpha_{sk} = 0$ or known, $k \geq 2$ is a common assumption that works in most applications. This linear combination of functions of x_k is most general

- 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

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

Takeaways/Outline

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

Applications to Lasso (4-5):

- 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
- GLM extension essentially (with exponential family requirement although this is w/in definition of GLM) is just a revisited Taylor Series expansion with analogous work and results and an additional(?) constraint that sample size scales in $\Omega(s \log p)$ for sparsity $|S| =: s$.

Questions/Notes

- (pg. 11 (c)) How does the tolerance $\tau_{\mathcal{L}}$ related to unidentifiable components in a high-dim model (possibly discussed in sections 6-8 of v1 paper)?
- Understand considerations of restricted eigenvalue, restricted isometry property, and partial Riesz conditions in the context of convergence fo ℓ_2 loss functionin ℓ_1 (i.e. lasso) settings

  See https://www.stat.berkeley.edu/~binyu/ps/papers2010/RaskuttiWY10.pdf Raskutti 2010 paper on correlated Gaussian ensembles section 2.2.2

  RE is listed as least severe, so curious if the others are just of historical importance or the different analyses of convergence are across these properties
- From Raskutti 2010, Σ-Gaussian ensemble design matrices have sample covariance matrices $X^T X/n$ that satisfy the RE property if Σ satisfies the RE condition (and we have adequate sample size scaling). Does Σ necessarily complete the RE condition as a covariance matrix? Seems like a nontrivial assumption

  https://www.stat.berkeley.edu/~binyu/ps/papers2010/RaskuttiWY10.pdf

Set-Up/Background

Currently assuming $\overline{\mathcal{M}} = \mathcal{M}$ until otherwise noted, deferring to $\mathcal{M}$ when referencing the model subspace.

Recall $\hat{\Delta} = \hat{\theta}_{\lambda_n} - \theta^*$. Also recall that the first-order Taylor series of any loss function $\mathcal{L} : \mathbb{R}^n \mapsto \mathbb{R}$ about $\theta^* + \Delta$[2] with some error $\delta\mathcal{L}(\theta^*, \Delta)$ is

$$\mathcal{L}(\theta^* + \Delta) = \mathcal{L}(\theta^*) + \Delta^T \nabla \mathcal{L}(\theta^*) + \delta\mathcal{L}(\theta^*, \Delta)$$

Definition: A loss function satisfies the restricted storng convexity condition with curvature $\kappa_{\mathcal{L}}$ and tolerance $\tau_{\mathcal{L}}(\theta^*)$ if its first-order Taylor's series error satisfies the following

$$\delta\mathcal{L}(\theta^*, \Delta) \geq \kappa_{\mathcal{L}} ||\Delta||_2^2 - \tau_{\mathcal{L}}^2(\theta^*), \quad \forall \Delta \in \mathbb{S}$$

for some set $\mathbb{S}$. We see that for reasonable selection of λ_n, we can identify $\mathbb{S} = \mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}, \theta^*)$.

OLS with exact sparsity ($\theta^* \in \mathcal{M}$), ℓ_1 regularization penalty

$\mathcal{L}(\theta) = ||y - X\theta||_2^2, \mathcal{R}(w) = ||w||_1$, and $S \subseteq [p]$ indexes the known, non-zero elements of θ (with $S_C = [p] \backslash S$).

In this setting $\mathcal{L}(\theta) := \frac{1}{2n}||y - X\theta||_2^2$ and thus $\nabla \mathcal{L}(\theta) = -\frac{1}{n} X^T(y - X\theta), \nabla^2 \mathcal{L}(\theta) = \frac{1}{n} X^T X$.

1. Decomposability satisfied for ℓ_1 penalty

2. Subspace compatibility fully specified by known ℓ_2, ℓ_1 relationship such that

$$\Psi(M) := \sup_{u \in \mathcal{M}} \frac{\mathcal{R}(u)}{||u||_2} = \sup_{u \in S} \frac{||u||_1}{||u||_2} = \sqrt{|S|}$$

3. First-order Taylor series expansion of OLS loss yields $\delta\mathcal{L}(\theta^*, \Delta) = \frac{1}{n}\Delta^T X^T X \Delta = \frac{1}{n}||X\Delta||_2^2$. Want to uniformly lower-bound this remainder for a sensible/appropriate set of Δ.

4. For properly selected regulation paramter λ_n (by Lemma 1, bound to be shown later), we can identify our appropriate set as $\mathbb{C} := \{||\Delta_{S_c}||_1 \leq 3||\Delta_S||_1\}$, as by assumption $\theta^* \in \mathcal{M}_S$, which implies $\Pi_{\mathcal{M}_{S\perp}}(\theta^*) = 0 \Rightarrow ||\theta^*_{\mathcal{M}_{S\perp}}||_1 = 0$.

   Desired bound: $\frac{1}{n}||X\theta||_2^2 \geq \kappa_{\mathcal{L}}||\theta||_2^2, \forall \theta \in \mathbb{C}$

5. Assuming sub-Gaussian(σ^2) noise w, we apply known results to bound $\frac{||X\theta||_2^2}{n}$, thus meeting the RE/RSC property if 1) $n > 64(\kappa_2/\kappa_1)^2$ (so that our lower bound is valid on $||X\Delta||_2^2/n$) and 2) $\lambda_n \geq 2||\nabla \mathcal{L}(\theta^*)||_\infty$ (so that our set of Δ, $\mathbb{C}$, is valid) [3]

[2] Noting Δ is some general perturbation, no explicit relationship to $\hat{\Delta}$

[3] Note that the sub-Gaussian result truly gives us a bound $\frac{||X\theta||_2}{\sqrt{n}} \geq \kappa_1||\theta||_2 - \kappa_2\sqrt{\frac{\log p}{n}}||\theta||_1$. The result that $||\theta||_1 \leq 4\sqrt{s}||\theta_M||_2$ from $\theta^* \in \mathcal{M}$ and clever constant selectionas mentioned allows us to extend this to the necessary bound $||X\theta||_2 \geq \kappa_{\mathcal{L}}||\theta||_2$.

6. We finish by identifying a satisfactory bound on λ_n such that previous analysis, assuming $\hat{\delta} \in \mathbb{C}$ is valid, namely $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$. Recall that we assume $w \sim subG(\sigma^2)$, where w is our random noise $w = Y - X\theta^*$.

$$\begin{aligned} 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*)) &= 2||\nabla\mathcal{L}(\theta^*)||_\infty \\ &= \frac{2}{n}||X^T(y - X\theta^*)||_\infty \\ &= \frac{2}{n}||X^T w||_\infty \end{aligned}$$

Recall that WLOG X is column-centered such that $\frac{||X_j||_2}{\sqrt{n}} \leq 1$:

$$\begin{aligned} \mathbb{P}\left(||X^T w||_\infty \geq t\right) &\leq \mathbb{P}\left(\sum_{j=1}^{p} |\langle X_j, w\rangle| \geq t\right) \\ &\leq 2p \exp\left\{-\frac{nt^2}{2\sigma^2}\right\} = 2 \exp\left\{-\frac{nt^2}{2\sigma^2} + \log p\right\} \end{aligned}$$

Informing λ_n selection such that above holds with high probability with respect to n, σ^2, and constants.

OLS with partial sparsity, ℓ_1 penalty ($\mathbb{B}_q$ setting)

Notation, $s \equiv |S|$.

We now no longer have (or assume) $\theta^* \in \mathcal{M}$, as we are approximating a weakly sparse vector with true sparsity. Thus $\mathcal{R}(\theta^*_{\mathcal{M}_\perp}) \neq 0$ (as $\Pi_{\mathcal{M}_\perp}(\theta^*) \neq 0$). Our results from the exact sparisy setting holds that for $\lambda_n = 4\sqrt{\frac{\sigma^2 \log p}{n}}$, $\lambda_n \geq ||\nabla\mathcal{L}(\theta^*)||_1$ with probability $1 - c_1 \exp\left\{-c_2 n \lambda_n^2\right\}$ (along with the subspace compatibility and decomposability arguments). It then only remains to demonstrate the RSC condition holds.

Now we cannot conclude $||\theta||_1 \leq 4\sqrt{s}||\theta_{\mathcal{M}}||_2$, as the space $\mathbb{C}$ is not a cone where $||\Delta_{\mathcal{M}}|| \leq 3||\Delta_{\mathcal{M}_\perp}||$ but now a star-shaped set (so $||\Delta||_1 \leq 4\sqrt{s}||\Delta||_2$ does not hold). The essence of this proof is now a control over the radius of the $\mathbb{B}_q(R_q)$ ball that we use to intersect $\mathbb{C}$.

GLM with Logit Link, ℓ_1 penalty

The Taylor Series approximation in the OLS setting is no longer exact such that $\delta\mathcal{L}(\theta^*, \Delta) \neq \Delta^T\nabla\mathcal{L}(\theta^*)$. We will appeal to a similar sub-Gaussian argument but must first more carefully bound $\delta\mathcal{L}$ and parameterize our design matrix (see GLM1 assumptions, bounding $\lambda_{min} \geq \kappa_\ell > 0$ of $cov(x_i)$ and sub-Gaussianity of $v^T x_i$, $\forall v \in \mathbb{R}$ with sub-Guassianity parameter $\sigma^2 \leq \kappa_u ||v||_2^2$).

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

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)

  Similarly, why is the Taylor series of the OLS loss function exact such that $\delta\mathcal{L}(\theta^*, \Delta) = \langle \Delta, \frac{1}{n}X^T X\Delta\rangle$? How is the hessian form introduced?
- End of page 12, how does this correspond to a restricted eigenvalue condition?

Proof of equivalent dual-norm definitions

WTS $\sup_{u\in\mathbb{R}^p\setminus\{0\}} \frac{\langle u,v\rangle}{\mathcal{R}(u)} = \sup_{\mathcal{R}(u)\le 1}\langle u,v\rangle$ (i.e. equivalence of two definitions of dual norm $R^*(u)$)

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

To demonstrate equivalence we prove the inequality in the opposite direction:

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

Prove ℓ_∞ is dual-norm of ℓ_1

Let $u, v \in \mathbb{R}^n$ and $v_{(n)} = \max_i |v_i|, i \in [n]$

$$\sup_{||u||\le 1} \langle u,v\rangle \le \sup_{||u||\le 1} \sum_i |u_i||v_i| \le \sup_{||u||\le 1} v_{(n)} \sum_i |u_i| \le ||v||_\infty$$

To demonstrate equivalence we prove the inequality in the opposite direction. Assume $v_j = ||v||_\infty$ (i.e. j indexes the maximum value in v), let $s \in \mathbb{R}^n, s_j = \text{sign}(v_j)$ and 0 elsewhere. Clearly $||s|| \in \{u : ||u|| \le 1\}$ and

$$\sup_{||u||\le 1} \langle u,v\rangle \ge \langle s,v\rangle = v_j = ||v||_\infty$$

Dual-norm of group-structured norm [4]

See 2.2 Example 2 (pg 5) for more details on group-sparsity

Consider $\theta \in \mathbb{R}^p$ such that $[p]$ is partitionable as $\mathcal{G} = \{G_1, G_2, ..., G_{N_\mathcal{G}}\}$ (i.e. $\mathcal{G}$ indexes a specific partinioning of $[p]$ of length $N_\mathcal{G}$). Consider similarly the vector $\overrightarrow{\alpha} = (\alpha_1, ..., \alpha_{N_\mathcal{G}}), \alpha_i \in [1,\infty]$.
We define the group norm of θ as

$$||\theta||_{\mathcal{G},\overrightarrow{\alpha}} := \sum_{t=1}^{N_\mathcal{G}} ||\theta_{G_t}||_{\alpha_t}$$

That is, for each tth partition, we take the α_t-norm of this partition of θ.

We want to identify the dual norm of $||\cdot||_{\mathcal{G},\overrightarrow{\alpha}}$ (limited to $\overrightarrow{\alpha} \in [2,\infty]^{N_\mathcal{G}}$). For convenience we drop the $\mathcal{G}$ index for the specific partitioning for the group norm.

[4]Forgoing some of Sahand's nicer/more concise notation for clarity/confirmation of understanding on my end

Define $\overrightarrow{\alpha^*}$ such that $\frac{1}{\alpha_t} + \frac{1}{\alpha_t^*} = 1$.

$$\begin{aligned}\sup_{||u||_{\overrightarrow{\alpha}} \leq 1} \langle u, v \rangle &= \sup_{||u||_{\overrightarrow{\alpha}} \leq 1} \sum_t \langle u_{G_t}, v_{G_t} \rangle \overset{\text{Hölder's}}{\leq} \sup_{||u||_{\overrightarrow{\alpha}} \leq 1} \sum_t ||u_{G_t}||_{\alpha_t} ||v_{G_t}||_{\alpha_t^*} \\ &\leq \sup_{||u||_{\overrightarrow{\alpha}} \leq 1} \max_w ||v_{G_w}||_{\alpha_w^*} \sum_t ||u_{G_t}||_{\alpha_t} = \max_w ||v_{G_w}||_{\alpha_w^*}\end{aligned}$$

Similarly define $k := \text{argmax}_w\, ||v_{G_w}||_{\alpha_w^*}$ and s such that for $k \in p$, $||s_{G_s}||_{\alpha_s} = 1$ and $s_i = 0, \forall i \notin G_s$. Clearly again $||s||_{\overrightarrow{\alpha}} = 1$:

$$\sup_{||u||_{\overrightarrow{\alpha}} \leq 1} \langle u, v \rangle \geq \langle s, v \rangle = \max_w ||v_{G_w}||_{\alpha_w^*} \quad \square \tag{1}$$

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[5] $\mathcal{C}$ 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 $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $\theta_{ij} = 0 \Leftrightarrow X_i \perp X_j |\text{rest}$

[5] 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

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

Exercises

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:

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

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

   (b)

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

   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) [6] 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:

[6] 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 $\{X_3, X_4, ..., X_p\} = X^* = X_{3,...,p}$:

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

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:

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

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

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 $\mathbf{S} = (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T$:

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

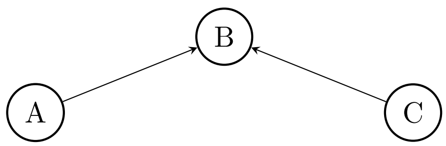
$$\begin{aligned}
\ell(\boldsymbol{\Theta}) &= \sum_{i=1}^{N} \log \left[(2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right] \\
&= \sum_{i=1}^{N} \log \left[(2\pi)^{-k/2}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

All of Statistics (Wasserman)

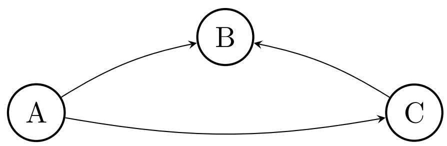
Chapter 17: Directed Graphs

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** (or Local Markov property) 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:

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

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

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

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

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

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

Simulations 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

$$\text{A)}\ \ X_1 - X_2 - X_3 \qquad \text{B)}\ \ X_1 \quad X_2 - X_3 \qquad \text{C)}\ \ X_1 \quad X_2 \quad X_3$$

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

2

$$\text{A)}\quad X_1 - X_2 - X_3 - X_4$$

$$\text{B)}\quad \begin{array}{c} X_1 - X_4 - X_2 \\ | \\ X_3 \end{array}$$

$$\text{C)}\quad X_2 - X_3 - X_4 - X_1$$

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*