Part III — Modern Statistical Methods

Theorems with proof

Based on lectures by R. D. Shah

Notes taken by Dexter Chua

Michaelmas 2017

The remarkable development of computing power and other technology now allows scientists and businesses to routinely collect datasets of immense size and complexity. Most classical statistical methods were designed for situations with many observations and a few, carefully chosen variables. However, we now often gather data where we have huge numbers of variables, in an attempt to capture as much information as we can about anything which might conceivably have an influence on the phenomenon of interest. This dramatic increase in the number variables makes modern datasets strikingly different, as well-established traditional methods perform either very poorly, or often do not work at all.

Developing methods that are able to extract meaningful information from these large and challenging datasets has recently been an area of intense research in statistics, machine learning and computer science. In this course, we will study some of the methods that have been developed to analyse such datasets. We aim to cover some of the following topics.

- Kernel machines: the kernel trick, the representer theorem, support vector machines, the hashing trick.
- Penalised regression: Ridge regression, the Lasso and variants.
- Graphical modelling: neighbourhood selection and the graphical Lasso. Causal inference through structural equation modelling; the PC algorithm.
- High-dimensional inference: the closed testing procedure and the Benjamini–Hochberg procedure; the debiased Lasso

Pre-requisites

Basic knowledge of statistics, probability, linear algebra and real analysis. Some background in optimisation would be helpful but is not essential.

Contents

0 Introduction

1 Classical statistics

Theorem (Cramér–Rao bound)**.** If $\tilde{\theta}$ is an unbiased estimator for θ, then $\mathrm{var}(\tilde{\theta}) - I^{-1}(\theta)$ is positive semi-definite.

Moreover, asymptotically, as $n \to \infty$, the maximum likelihood estimator is unbiased and achieves the Carmér–Rao bound.

2 Kernel machines

2.1 Ridge regression

Theorem. Suppose $\text{rank}(X) = p$. Then for $\lambda > 0$ sufficiently small (depending on β^0 and σ^2), the matrix

$$\mathbb{E}(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T - \mathbb{E}(\hat{\beta}^R_\lambda - \beta^0)(\hat{\beta}^R_\lambda - \beta^0)^T \qquad (*)$$

is positive definite.

Proof. We know that the first term is just $\sigma^2(X^TX)^{-1}$. The right-hand-side has a variance term and a bias term. We first look at the bias:

$$\begin{aligned}\mathbb{E}[\hat{\beta} - \beta^0] &= (X^TX + \lambda I)^{-1}X^TX\beta^0 - \beta^0\\ &= (X^TX + \lambda I)^{-1}(X^TX + \lambda I - \lambda I)\beta^0 - \beta^0\\ &= -\lambda(X^TX + \lambda I)^{-1}\beta^0.\end{aligned}$$

We can also compute the variance

$$\text{var}(\hat{\beta}) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}.$$

Note that both terms appearing in the squared error look like

$$(X^TX + \lambda I)^{-1}\text{something}(X^TX + \lambda I)^{-1}.$$

So let's try to write $\sigma^2(X^TX)^{-1}$ in this form. Note that

$$\begin{aligned}(X^TX)^{-1} &= (X^TX + \lambda I)^{-1}(X^TX + \lambda I)(X^TX)^{-1}(X^TX + \lambda I)(X^TX + \lambda I)^{-1}\\ &= (X^TX + \lambda I)^{-1}(X^TX + 2\lambda I + \lambda^2(X^TX)^{-1})(X^TX + \lambda I)^{-1}.\end{aligned}$$

Thus, we can write $(*)$ as

$$\begin{aligned}&(X^TX + \lambda I)^{-1}\Big(\sigma^2(X^TX + 2\lambda I + \lambda^2(X^TX)^{-1})\\ &\qquad - \sigma^2X^TX - \lambda^2\beta^0(\beta^0)^T\Big)(X^TX + \lambda I)^{-1}\\ &= \lambda(X^TX + \lambda I)^{-1}\Big(2\sigma^2 I + \lambda(X^TX)^{-1} - \lambda\beta^0(\beta^0)^T\Big)(X^TX + \lambda I)^{-1}\end{aligned}$$

Since $\lambda > 0$, this is positive definite iff

$$2\sigma^2 I + \sigma^2\lambda(X^TX)^{-1} - \lambda\beta^0(\beta^0)^T$$

is positive definite, which is true for $0 < \lambda < \frac{2\sigma^2}{\|\beta^0\|_2^2}$. □

Theorem (Singular value decomposition)**.** Let $X \in \mathbb{R}^{n\times p}$ be any matrix. Then it has a *singular value decomposition* (SVD)

$$\underset{n\times p}{X} = \underset{n\times n}{U}\ \underset{n\times p}{D}\ \underset{p\times p}{V^T},$$

where U, V are orthogonal matrices, and $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$, where $m = \min(n, p)$, and all other entries of D are zero.

2.2 v-fold cross-validation

2.3 The kernel trick

Proposition. Given $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{H}$, define $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Then for any $x_1, \dots, x_n \in \mathcal{X}$, the matrix $K \in \mathbb{R}^n \times \mathbb{R}^n$ with entries

$$K_{ij} = k(x_i, x_j)$$

is positive semi-definite.

Proof. Let $x_1, \dots, x_n \in \mathcal{X}$, and $\alpha \in \mathbb{R}^n$. Then

$$\begin{aligned} \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle \\ &\geq 0 \end{aligned}$$

since the inner product is positive definite. □

Theorem (Moore–Aronszajn theorem)**.** For every kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists an inner product space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Proof. Let $\mathcal{H}$ denote the vector space of functions $f : \mathcal{X} \to \mathbb{R}$ of the form

$$f(\,\cdot\,) = \sum_{i=1}^{n} \alpha_i k(\,\cdot\,, x_i) \tag{$*$}$$

for some $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$. If

$$g(\,\cdot\,) = \sum_{i=1}^{m} \beta_j k(\,\cdot\,, x'_j) \in \mathcal{H},$$

then we tentatively define the inner product of f and g by

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j).$$

We have to check that this is an inner product, but even before that, we need to check that this is well-defined, since f and g can be represented in the form $(*)$ in multiple ways. To do so, simply observe that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(x'_j). \tag{$\dagger$}$$

The first equality shows that the definition of our inner product does not depend on the representation of g, while the second equality shows that it doesn't depend on the representation of f.

To show this is an inner product, note that it is clearly symmetric and bilinear. To show it is positive definite, note that we have

$$\langle f, f\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i k(x_i, x_j)\alpha_j \geq 0$$

since the kernel is positive semi-definite. It remains to check that if $\langle f, f\rangle = 0$, then $f = 0$ as a function. To this end, note the important *reproducing property*: by (†), we have

$$\langle k(\,\cdot\,, x), f\rangle = f(x).$$

This says $k(\cdot, x)$ represents the evaluation-at-x linear functional.

In particular, we have

$$f(x)^2 = \langle k(\,\cdot\,, x), f\rangle^2 \leq \langle k(\,\cdot\,, x), k(\,\cdot\,, x)\rangle\langle f, f\rangle = 0.$$

Here we used the Cauchy–Schwarz inequality, which, if you inspect the proof, does not require positive definiteness, just positive semi-definiteness. So it follows that $f \equiv 0$. Thus, we know that $\mathcal{H}$ is indeed an inner product space.

We now have to construct a feature map. Define $\phi : \mathcal{X} \to \mathcal{H}$ by

$$\phi(x) = k(\,\cdot\,, x).$$

Then we have already observed that

$$\langle \phi(x), \phi(x')\rangle = \langle k(\,\cdot\,, x), k(\,\cdot\,, x')\rangle = k(x, x'),$$

as desired. □

2.4 Making predictions

Theorem (Representer theorem)**.** Let $\mathcal{H}$ be an RKHS with reproducing kernel k. Let c be an arbitrary loss function and $J : [0, \infty) \to \mathbb{R}$ any strictly increasing function. Then the minimizer $\hat{f} \in \mathcal{H}$ of

$$Q_1(f) = c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + J(\|f\|_{\mathcal{H}}^2)$$

lies in the linear span of $\{k(\,\cdot\,, x_i)\}_{i=1}^n$.

Proof. Suppose $\hat{f}$ minimizes Q_1. We can then write

$$\hat{f} = u + v$$

where $u \in V = \operatorname{span}\{k(\,\cdot\,, x_i) : i = 1, \ldots, n\}$ and $v \in V^\perp$. Then

$$\hat{f}(x_i) = \langle f, k(\,\cdot\,, x_i)\rangle = \langle u + v, k(\,\cdot\,, x_i)\rangle = \langle u, k(\,\cdot\,, x_i)\rangle = u(x_i).$$

So we know that

$$c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) = c(Y, x_1, \ldots, x_n, u(x_1), \ldots, u(x_n)).$$

Meanwhile,

$$\|f\|_{\mathcal{H}}^2 = \|u+v\|_{\mathcal{H}}^2 = \|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2,$$

using the fact that u and v are orthogonal. So we know

$$J(\|f\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$$

with equality iff $v = 0$. Hence $Q_1(f) \geq Q_1(u)$ with equality iff $v = 0$, and so we must have $v = 0$ by optimality.

Thus, we know that the optimizer in fact lies in V, and then the formula of Q_2 just expresses Q_1 in terms of α. □

Theorem. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}(f^0(x_i) - \hat{f}_\lambda(x_i))^2 &\leq \frac{\sigma^2}{n}\sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2} + \frac{\lambda}{4n}\\ &\leq \frac{\sigma^2}{n}\frac{1}{\lambda}\sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda\right) + \frac{\lambda}{4n}.\end{aligned}$$

Proof. We know from the representer theorem that

$$(\hat{f}_\lambda(x_i), \dots, \hat{f}_\lambda(x_n))^T = K(K+\lambda I)^{-1}Y.$$

Also, there is some $\alpha \in \mathbb{R}^n$ such that

$$(f^0(x_1), \dots, f^0(x_n))^T = K\alpha.$$

Moreover, on the example sheet, we show that

$$1 \geq \|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha.$$

Consider the eigen-decomposition $K = UDU^T$, where U is orthogonal, $D_{ii} = d_i$ and $D_{ij} = 0$ for $i \neq j$. Then we have

$$\mathbb{E}\sum_{i=1}^n (f^0(x_i) - \hat{f}_\lambda(x_i))^2 = \mathbb{E}\|K\alpha - K(K+\lambda I)^{-1}(K\alpha + \varepsilon)\|_2^2$$

Noting that $K\alpha = (K+\lambda I)(K+\lambda I)^{-1}K\alpha$, we obtain

$$\begin{aligned}&= \mathbb{E}\left\|\lambda(K+\lambda I)^{-1}K\alpha - K(K+\lambda I)^{-1}\varepsilon\right\|_2^2\\ &= \underbrace{\lambda^2\left\|(K+\lambda I)^{-1}K\alpha\right\|_2^2}_{\text{(I)}} + \underbrace{\mathbb{E}\left\|K(K+\lambda I)^{-1}\varepsilon\right\|_2^2}_{\text{(II)}}.\end{aligned}$$

At this stage, we can throw in the eigen-decomposition of K to write (I) as

$$\begin{aligned}\text{(I)} &= \lambda^2\left\|(UDU^T + \lambda I)^{-1}UDU^T\alpha\right\|_2^2\\ &= \lambda^2\|U(D+\lambda I)^{-1}\underbrace{DU^T\alpha}_{\theta}\|_2^2\\ &= \sum_{i=1}^n \theta_i^2 \frac{\lambda^2}{(d_i+\lambda)^2}\end{aligned}$$

Now we have

$$\alpha^T K\alpha = \alpha^T UDU^T\alpha = \alpha^T UDD^+DU^T,$$

where D^+ is diagonal and

$$D^+_{ii} = \begin{cases} d_i^{-1} & d_i > 0 \\ 0 & \text{otherwise} \end{cases}.$$

We can then write this as

$$\alpha^T K\alpha = \sum_{d_i>0}\frac{\theta_i^2}{d_i}.$$

The key thing we know about this is that it is ≤ 1.

Note that by definition of θ_i, we see that if $d_i = 0$, then $\theta_i = 0$. So we can write

$$(\text{II}) = \sum_{i:d_i\geq 0}\frac{\theta_i^2}{d_i}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \leq \lambda \max_{i=1,\dots,n}\frac{d_i\lambda}{(d_i+\lambda)^2}$$

by Hölder's inequality with $(p, q) = (1, \infty)$. Finally, use the inequality that

$$(a+b)^2 \geq 4ab$$

to see that we have

$$(\text{I}) \leq \frac{\lambda}{4}.$$

So we are left with (II), which is a bit easier. Using the trace trick, we have

$$\begin{aligned}
(\text{II}) &= \mathbb{E}\varepsilon^T(K+\lambda I)^{-1}K^2(K+\lambda I)^{-1}\varepsilon \\
&= \mathbb{E}\operatorname{tr}\left(K(K+\lambda I)^{-1}\varepsilon\varepsilon^T(K+\lambda I)^{-1}K\right) \\
&= \operatorname{tr}\left(K(K+\lambda I)^{-1}\mathbb{E}(\varepsilon\varepsilon^T)(K+\lambda I)^{-1}K\right) \\
&= \sigma^2\operatorname{tr}\left(K^2(K+\lambda I)^{-2}\right) \\
&= \sigma^2\operatorname{tr}\left(UD^2U^T(UDU^T+\lambda I)^{-2}\right) \\
&= \sigma^2\operatorname{tr}\left(D^2(D+\lambda I)^{-2}\right) \\
&= \sigma^2\sum_{i=1}^n\frac{d_i^2}{(d_i+\lambda)^2}.
\end{aligned}$$

Finally, writing $\frac{d_i^2}{(d_i+\lambda)^2} = \frac{d_i}{\lambda}\frac{d_i\lambda}{(d_i+\lambda)^2}$, we have

$$\frac{d_i^2}{(d_i+\lambda)^2} \leq \min\left(1, \frac{d_i}{4\lambda}\right),$$

and we have the second bound. □

2.5 Other kernel machines

2.6 Large-scale kernel machines

Theorem (Bochner's theorem)**.** Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant if and only if there exists some distribution F on $\mathbb{R}^p$ and $c > 0$ such that if $W \sim F$, then

$$k(x, x') = c\mathbb{E}\cos((x-x')^TW).$$

3 The Lasso and beyond

3.1 The Lasso estimator

Theorem. Let $\hat{\beta}$ be the Lasso solution with

$$\lambda = A\sigma\sqrt{\frac{\log p}{n}}$$

for some A. Then with probability $1 - 2p^{-(A^2/2-1)}$, we have

$$\frac{1}{n}\|X\beta^0 - X\hat{\beta}\|_2^2 \leq 4A\sigma\sqrt{\frac{\log p}{n}}\|\beta^0\|_1.$$

Proof. We don't really have a closed form solution for $\hat{\beta}$, and in general it doesn't exist. So the only thing we can use is that it in fact minimizes $Q_\lambda(\beta)$. Thus, by definition, we have

$$\frac{1}{2n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^0\|_2^2 + \lambda\|\beta^0\|_1.$$

We know exactly what Y is. It is $X\beta^0 + \varepsilon - \bar{\varepsilon}1$. If we plug this in, we get

$$\frac{1}{2n}\|X\beta^0 - X\hat{\beta}\|_2^2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Here we use the fact that X is centered, and so is orthogonal to 1.

Now Hölder tells us

$$|\varepsilon^T X(\hat{\beta} - \beta^0)| \leq \|X^T\varepsilon\|_\infty\|\hat{\beta} - \beta^0\|_1.$$

We'd like to bound $\|X^T\varepsilon\|_\infty$, but it can be arbitrarily large since ε is a Gaussian. However, with high probability, it is small. Precisely, define

$$\Omega = \left\{\frac{1}{n}\|X^T\varepsilon\|_\infty \leq \lambda\right\}.$$

In a later lemma, we will show later that $\mathbb{P}(\Omega) \geq 1 - 2p^{-(A^2/2-1)}$. Assuming Ω holds, we have

$$\frac{1}{2n}\|X\beta^0 - X\hat{\beta}\|_2^2 \leq \lambda\|\hat{\beta} - \beta^0\| - \lambda\|\hat{\beta}\| + \lambda\|\beta^0\| \leq 2\lambda\|\beta^0\|_1. \qquad \square$$

3.2 Basic concentration inequalities

Lemma (Markov's inequality)**.** Let W be a non-negative random variable. Then

$$\mathbb{P}(W \geq t) \leq \frac{1}{t}\mathbb{E}W.$$

Proof. We have

$$t\mathbf{1}_{W \geq t} \leq W.$$

The result then follows from taking expectations and then dividing both sides by t. $\square$

Corollary (Chernoff bound)**.** For any random variable W, we have

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t} \mathbb{E} e^{\alpha W}.$$

Corollary. Any sub-Gaussian random variable W with parameter σ satisfies

$$\mathbb{P}(W \geq t) \leq e^{-t^2/2\sigma^2}. \qquad \square$$

Lemma (Hoeffding's lemma)**.** If W has mean zero and takes values in $[a, b]$, then W is sub-Gaussian with parameter $\frac{b-a}{2}$. $\square$

Proposition. Let $(W_i)_{i=1}^n$ be independent mean-zero sub-Gaussian random variables with parameters $(\sigma_i)_{i=0}^n$, and let $\gamma \in \mathbb{R}^n$. Then $\gamma^T W$ is sub-Gaussian with parameter

$$\left(\sum (\gamma_i \sigma_i)^2\right)^{1/2}.$$

Proof. We have

$$\begin{aligned}
\mathbb{E} \exp\left(\alpha \sum_{i=1}^n \gamma_i W_i\right) &= \prod_{i=1}^n \mathbb{E} \exp\left(\alpha \gamma_i W_i\right) \\
&\leq \prod_{i=1}^n \exp\left(\frac{\alpha^2}{2} \gamma_i^2 \sigma_i^2\right) \\
&= \exp\left(\frac{\alpha^2}{2} \sum_{i=1}^n \sigma_i^2 \gamma_i^2\right). \qquad \square
\end{aligned}$$

Lemma. Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero sub-Gaussian with common parameter σ. Let

$$\lambda = A\sigma \sqrt{\frac{\log p}{n}}.$$

Let X be a matrix whose columns all have norm $\sqrt{n}$. Then

$$\mathbb{P}\left(\frac{1}{n} \|X^T \varepsilon\|_\infty \leq \lambda\right) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proof. We have

$$\mathbb{P}\left(\frac{1}{n} \|X^T \varepsilon\|_\infty > \lambda\right) \leq \sum_{j=1}^p \mathbb{P}\left(\frac{1}{n} |X_j^T \varepsilon| > \lambda\right).$$

But $\pm \frac{1}{n} X_j^T \varepsilon$ are both sub-Gaussian with parameter

$$\sigma \left(\sum_i \left(\frac{X_{ij}}{n}\right)^2\right)^{1/2} = \frac{\sigma}{\sqrt{n}}.$$

Then by our previous corollary, we get

$$\sum_{j=1}^p \mathbb{P}\left(\frac{1}{n} |X_j^T \varepsilon|_\infty > \lambda\right) \leq 2p \exp\left(-\frac{\lambda^2 n}{2\sigma^2}\right).$$

Note that we have the factor of 2 since we need to consider the two cases $\frac{1}{n}X_j^T\varepsilon > \lambda$ and $-\frac{1}{n}X_j^T\varepsilon > \lambda$.

Plugging in our expression of λ, we write the bound as

$$2p\exp\left(-\frac{1}{2}A^2\log p\right) = 2p^{1-A^2/2}. \qquad \square$$

Proposition (Bernstein's inequality)**.** Let $W_1, W_2, \ldots, W_n$ be independent random variables with $\mathbb{E}W_i = \mu$, and suppose each W_i satisfies Bernstein's condition with parameters (σ, b). Then

$$\mathbb{E}e^{\alpha(W_i-\mu)} \le \exp\left(\frac{\alpha^2\sigma^2/2}{1-b|\alpha|}\right) \text{ for all } |\alpha| < \frac{1}{b},$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n W_i - \mu \ge t\right) \le \exp\left(-\frac{nt^2}{2(\sigma^2+bt)}\right) \text{ for all } t \ge 0.$$

Proof. For the first part, we fix i and write $W = W_i$. Let $|\alpha| < \frac{1}{b}$. Then

$$\begin{aligned}
\mathbb{E}e^{\alpha(W_i-\mu)} &= \sum_{k=0}^\infty \mathbb{E}\left[\frac{1}{k!}\alpha^k|W_i-\mu|^k\right] \\
&\le 1 + \frac{\sigma^2\alpha^2}{2}\sum_{k=2}^\infty |\alpha|^{k-2}b^{k-2} \\
&= 1 + \frac{\sigma^2\alpha^2}{2}\frac{1}{1-|\alpha|b} \\
&\le \exp\left(\frac{\alpha^2\sigma^2/2}{1-b|\alpha|}\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\exp\left(\frac{1}{n}\sum_{i=1}^n \alpha(W_i-\mu)\right) &= \prod_{i=1}^n \mathbb{E}\exp\left(\frac{\alpha}{n}(W_i-\mu)\right) \\
&\le \exp\left(n\frac{\left(\frac{\alpha}{n}\right)^2\sigma^2/2}{1-b\left|\frac{\alpha}{n}\right|}\right),
\end{aligned}$$

assuming $\left|\frac{\alpha}{n}\right| < \frac{1}{b}$. So it follows that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n W_i - \mu \ge t\right) \le e^{-\alpha t}\exp\left(n\frac{\left(\frac{\alpha}{n}\right)^2\sigma^2/2}{1-b\left|\frac{\alpha}{n}\right|}\right).$$

Setting

$$\frac{\alpha}{n} = \frac{t}{bt+\sigma^2} \in \left[0, \frac{1}{b}\right)$$

gives the result. $\square$

Lemma. Let W, Z be mean-zero sub-Gaussian random variables with parameters σ_W and σ_Z respectively. Then WZ satisfies Bernstein's condition with parameter $(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$.

Proof. For any random variable Y (which we will later take to be WZ), for $k > 1$, we know

$$\begin{aligned}\mathbb{E}|Y - \mathbb{E}Y|^k &= 2^k \mathbb{E}\left|\frac{1}{2}Y - \frac{1}{2}\mathbb{E}Y\right|^k \\ &\leq 2^k \mathbb{E}\left|\frac{1}{2}|Y| + \frac{1}{2}|\mathbb{E}Y|\right|^k.\end{aligned}$$

Note that

$$\left|\frac{1}{2}|Y| + \frac{1}{2}|\mathbb{E}Y|\right|^k \leq \frac{|Y|^k + |\mathbb{E}Y|^k}{2}$$

by Jensen's inequality. Applying Jensen's inequality again, we have

$$|\mathbb{E}Y|^k \leq \mathbb{E}|Y|^k.$$

Putting the whole thing together, we have

$$\mathbb{E}|Y - \mathbb{E}Y|^k \leq 2^k \mathbb{E}|Y|^k.$$

Now take $Y = WZ$. Then

$$\mathbb{E}|WZ - \mathbb{E}WZ| \leq 2^k \mathbb{E}|WZ|^k \leq 2^k \left(\mathbb{E}W^{2k}\right)^{1/2} (\mathbb{E}Z^{2k})^{1/2},$$

by the Cauchy–Schwarz inequality.

We know that sub-Gaussians satisfy a bound on the tail probability. We can then use this to bound the moments of W and Z. First note that

$$W^{2k} = \int_0^\infty \mathbf{1}_{x<W^{2k}}\, \mathrm{d}x.$$

Taking expectations of both sides, we get

$$\mathbb{E}W^{2k} = \int_0^\infty \mathbb{P}(x < W^{2k})\, \mathrm{d}x.$$

Since we have a tail bound on W instead of W^{2k}, we substitute $x = t^{2k}$. Then $\mathrm{d}x = 2kt^{2k-1}\, \mathrm{d}t$. So we get

$$\begin{aligned}\mathbb{E}W^{2k} &= 2k \int_0^\infty t^{2k-1} \mathbb{P}(|W| > t)\, \mathrm{d}t \\ &= 4k \int_0^\infty t^{2k} \exp\left(-\frac{t^2}{2\sigma_N^2}\right)\, \mathrm{d}t.\end{aligned}$$

where again we have a factor of 2 to account for both signs. We perform yet another substitution

$$x = \frac{t^2}{2\sigma_N^2}, \quad \mathrm{d}x = \frac{t}{\sigma_W^2}\, \mathrm{d}t.$$

Then we get

$$\mathbb{E}W^{2k} = 2^{k+1}\sigma_W^{2k} k \sigma_W^2 \int_0^\infty x^{k-1} e^{-x}\, \mathrm{d}x = 4 \cdot k! \sigma_W^2.$$

Plugging this back in, we have

$$\begin{aligned}\mathbb{E}|WZ - \mathbb{E}WZ|^k &\le 2^k 2^{k+1} k! \sigma_W^k \sigma_Z^k \sigma_Z^k \\ &= \frac{1}{2} k! 2^{2k+2} \sigma_W^k \sigma_Z^k \\ &= \frac{1}{2} k! (8\sigma_W \sigma_Z)^2 (4\sigma_W \sigma_Z)^{k-2}. \qquad \square\end{aligned}$$

3.3 Convex analysis and optimization theory

Proposition.

(i) Let $f_1, \dots, f_m : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex with $\operatorname{dom} f_1 \cap \cdots \cap \operatorname{dom} f_m \neq \emptyset$, and let $c_1, \dots, c_m \ge 0$. Then $c_1 + \cdots + c_m f_m$ is a convex function.

(ii) If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then

  (a) f is convex iff its Hessian is positive semi-definite everywhere.

  (b) f is strictly convex if its Hessian positive definite everywhere. $\square$

Proposition. Let f be convex and differentiable at $x \in \operatorname{int}(\operatorname{dom} f)$. Then $\partial f(x) = \{\nabla f(x)\}$. $\square$

Proposition. Suppose f and g are convex with $\operatorname{int}(\operatorname{dom} f) \cap \operatorname{int}(\operatorname{dom} g) \neq \emptyset$, and $\alpha > 0$. Then

$$\begin{aligned}\partial(\alpha f)(x) &= \alpha \partial f(x) = \{\alpha v : v \in \partial f(x)\} \\ \partial(f + g)(x) &= \partial g(x) + \partial f(x). \qquad \square\end{aligned}$$

Proposition. If f is convex, then

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) \Leftrightarrow 0 \in \partial f(x^*).$$

Proof. Both sides are equivalent to the requirement that $f(y) \ge f(x^*)$ for all y. $\square$

Proposition. For $x \in \mathbb{R}^d$ and $A \in \{j : x_j \neq 0\}$, we have

$$\partial \|x\|_1 = \{v \in \mathbb{R}^d : \|v\|_\infty \le 1, v_A = \operatorname{sgn}(x_A)\}.$$

Proof. It suffices to look at the subdifferential of the absolute value function, and then add them up.

For $j = 1, \dots, d$, we define $g_j : \mathbb{R}^d \to \mathbb{R}$ by sending x to $|x_j|$. If $x_j \neq 0$, then g_j is differentiable at x, and so we know $\partial g_j(x) = \{\operatorname{sgn}(x_j) e_j\}$, with e_j the jth standard basis vector.

When $x_j = 0$, if $v \in \partial g_j(x_j)$, then

$$g_j(y) \ge g_j(x) + v^T (y - x).$$

So

$$|y_j| \ge v^T (y - x).$$

We claim this holds iff $v_j \in [-1, 1]$ and $v_{-j} = 0$. The $\Leftarrow$ direction is an immediate calculation, and to show $\Rightarrow$, we pick $y_{-j} = v_{-j} + x_{-j}$, and $y_j = 0$. Then we have

$$0 \geq v_{-j}^T v_{-j}.$$

So we know that $v_{-j} = 0$. Once we know this, the condition says

$$|y_j| \geq v_j y_j$$

for all y_j. This is then true iff $v_j \in [-1, 1]$. Forming the set sum of the $\partial g_j(x)$ gives the result. □

3.4 Properties of Lasso solutions

Proposition. $X\hat{\beta}_\lambda^L$ is unique.

Proof. Fix $\lambda > 0$ and stop writing it. Suppose $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ are two Lasso solutions at λ. Then

$$Q(\hat{\beta}^{(1)}) = Q(\hat{\beta}^{(2)}) = c^*.$$

As Q is convex, we know

$$c^* = Q\left(\frac{1}{2}\hat{\beta}^{(1)} + \frac{1}{2}\hat{\beta}^{(2)}\right) \leq \frac{1}{2}Q(\hat{\beta}^{(1)}) + \frac{1}{2}Q(\hat{\beta}^{(2)}) = c^*.$$

So $\frac{1}{2}\hat{\beta}^{(1)} + \frac{1}{2}\hat{\beta}^{(2)}$ is also a minimizer.

Since the two terms in $Q(\beta)$ are individually convex, it must be the case that

$$\left\|\frac{1}{2}(Y - X\hat{\beta}^{(1)}) + \frac{1}{2}(Y - X\hat{\beta}^{(2)})\right\|_2^2 = \frac{1}{2}\left\|Y - X\hat{\beta}^{(1)}\right\|_2^2 + \frac{1}{2}\left\|Y - X\hat{\beta}^{(2)}\right\|_2^2$$
$$\left\|\frac{1}{2}(\hat{\beta}^{(1)} + \hat{\beta}^{(2)})\right\|_1 = \frac{1}{2}\|\hat{\beta}^{(1)}\|_1 + \frac{1}{2}\|\hat{\beta}^{(2)}\|_1.$$

Moreover, since $\|\cdot\|_2^2$ is strictly convex, we can have equality only if $X\hat{\beta}^{(1)} = X\hat{\beta}^{(2)}$. So we are done. □

3.5 Variable selection

Theorem.

(i) If $\|\Delta\|_\infty \leq 1$, or equivalently

$$\max_{k \in N} |\operatorname{sgn}(\beta_S^0)^T (X_S^T X_S)^{-1} X_S^T X_k| \leq 1,$$

and moreover

$$|\beta_k^0| > \lambda \left|\operatorname{sgn}(\beta_S^0)^T \left(\frac{1}{n} X_j^T X_j\right)_k^{-1}\right|$$

for all $k \in S$, then there exists a Lasso solution $\hat{\beta}_\lambda^L$ with $\operatorname{sgn}(\hat{\beta}_\lambda^L) = \operatorname{sgn}(\beta^0)$.

(ii) If there exists a Lasso solution with $\operatorname{sgn}(\hat{\beta}_\lambda^L) = \operatorname{sgn}(\beta^0)$, then $\|\Delta\|_\infty \leq 1$.

Proof. Write $\hat{\beta} = \hat{\beta}^L_\lambda$, and write $\hat{S} = \{k : \hat{\beta}_k \neq 0\}$. Then the KKT conditions are that

$$\frac{1}{n} X^T(\beta^0 - \hat{\beta}) = \lambda \hat{\nu},$$

where $\|\hat{\nu}\|_\infty \leq 1$ and $\hat{\nu}_{\hat{S}} = \operatorname{sgn}(\hat{\beta}_{\hat{S}})$.

We expand this to say

$$\frac{1}{n}\begin{pmatrix} X_S^T X_S & X_S^T X_N \\ X_N^T X_S & X N^T X_N \end{pmatrix} \begin{pmatrix} \beta_S^0 - \hat{\beta}_S \\ -\hat{\beta}_N \end{pmatrix} = \lambda \begin{pmatrix} \hat{\nu}_S \\ \hat{\nu}_N \end{pmatrix}.$$

Call the top and bottom equations (1) and (2) respectively.

It is easier to prove (ii) first. If there is such a solution, then $\hat{\beta}_N = 0$. So from (1), we must have

$$\frac{1}{n} X_S^T X_S(\beta_S^0 - \hat{\beta}_S) = \lambda \hat{\nu}_S.$$

Inverting $\frac{1}{n} X_S^T X_S$, we learn that

$$\beta_S^0 - \hat{\beta}_S = \lambda \left(\frac{1}{n} X_S^T X_S\right)^{-1} \operatorname{sgn}(\beta_S^0).$$

Substitute this into (2) to get

$$\lambda \frac{1}{n} X_N^T X_S \left(\frac{1}{n} X_S^T X_S\right)^{-1} \operatorname{sgn}(\beta_S^0) = \lambda \hat{\nu}_N.$$

By the KKT conditions, we know $\|\hat{\nu}_N\|_\infty \leq 1$, and the LHS is exactly $\lambda \Delta$.

To prove (1), we need to exhibit a $\hat{\beta}$ that agrees in sign with $\hat{\beta}$ and satisfies the equations (1) and (2). In particular, $\hat{\beta}_N = 0$. So we try

$$\begin{aligned}(\hat{\beta}_S, \hat{\nu}_S) &= \left(\beta_S^0 - \lambda \left(\frac{1}{n} X_S^T X_S\right)^{-1} \operatorname{sgn}(\beta_S^0), \operatorname{sgn}(\beta_S^0)\right) \\ (\hat{\beta}_N, \nu_N) &= (0, \Delta).\end{aligned}$$

This is by construction a solution. We then only need to check that

$$\operatorname{sgn}(\hat{\beta}_S) = \operatorname{sgn}(\beta_S^0),$$

which follows from the second condition. □

Theorem. Assume $\phi^2 > 0$, and let $\hat{\beta}$ be the Lasso solution with

$$\lambda = A\sigma\sqrt{\log p / n}.$$

Then with probability at least $1 - 2p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda \|\hat{\beta} - \beta^0\|_1 \leq \frac{16\lambda^2 s}{\phi^2} = \frac{16A^2 \log p}{\phi^2} \frac{s\sigma^2}{n}.$$

Proof. Start with the basic inequality $Q_\lambda(\hat{\beta}) \leq Q_\lambda(\beta^0)$, which gives us

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1.$$

We work on the event

$$\Omega = \left\{\frac{1}{n}\|X^T\varepsilon\|_\infty \leq \frac{1}{2}\lambda\right\},$$

where after applying Hölder's inequality, we get

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - \beta^0\|_1 + 2\lambda\|\beta^0\|_1.$$

We can move $2\lambda\|\hat{\beta}\|_1$ to the other side, and applying the triangle inequality, we have

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 \leq 3\lambda\|\hat{\beta} - \beta^0\|.$$

If we manage to bound the RHS from above as well, so that

$$3\lambda\|\hat{\beta} - \beta^0\| \leq c\lambda\frac{1}{\sqrt{n}}\|X(\hat{\beta} - \beta^0)\|_2$$

for some c, then we obtain the bound

$$\frac{1}{n}\|X(\beta - \beta^0)\|_2^2 \leq c^2\lambda^2.$$

Plugging this back into the second bound, we also have

$$3\lambda\|\hat{\beta} - \beta^0\|_1 \leq c^2\lambda^2.$$

To obtain these bounds, we want to apply the definition of ϕ^2 to $\hat{\beta} - \beta^0$. We thus need to show that the $\hat{\beta} - \beta^0$ satisfies the conditions required in the infimum taken.

Write

$$a = \frac{1}{n\lambda}\|X(\hat{\beta} - \beta^0)\|_2^2.$$

Then we have

$$a + 2(\|\hat{\beta}_n\|_1 + \|\hat{\beta}_S\|_1) \leq \|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_N\|_1 + 2\|\beta_S^0\|_1.$$

Simplifying, we obtain

$$a + \|\hat{\beta}_N\|_1 \leq \|\hat{\beta}_S - \beta_S^0\|_1 + 2\|\beta_S^0\|_1 - 2\|\hat{\beta}_S\|_1.$$

Using the triangle inequality, we write this as

$$a + \|\hat{\beta}_N - \beta^0\|_N \leq 3\|\hat{\beta}_S - \beta_S^0\|_1.$$

So we immediately know we can apply the compatibility condition, which gives us

$$\phi^2 \leq \frac{\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2}{\frac{1}{s}\|\hat{\beta}_S - \beta_S^0\|_1^2}.$$

Also, we have

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}-\beta^0\|_1\leq 4\lambda\|\hat{\beta}_S-\beta^0_S\|_1.$$

Thus, using the compatibility condition, we have

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}-\beta^0\|\leq\frac{4\lambda}{\phi}\sqrt{\frac{s}{n}}\|X(\hat{\beta}-\beta^0)\|_2.$$

Thus, dividing through by $\frac{1}{\sqrt{n}}\|X(\hat{\beta}-\beta^0)\|_2$, we obtain

$$\frac{1}{\sqrt{n}}\|X(\hat{\beta}-\beta^0)\|_2\leq\frac{4\lambda\sqrt{s}}{\phi}. \qquad (*)$$

So we substitute into the RHS $(*)$ and obtain

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}-\beta^0\|_1\leq\frac{16\lambda^2 s}{\phi^2}. \qquad \square$$

Lemma. Let $\Theta,\Sigma\in\mathbb{R}^{p\times p}$. Suppose $\phi^2_\Theta(S)>0$ and

$$\max_{j,k}|\Theta_{jk}-\Sigma_{jk}|\leq\frac{\phi^2_\Theta(S)}{32|S|}.$$

Then

$$\phi^2_\Sigma(S)\geq\frac{1}{2}\phi^2_\Theta(S).$$

Proof. We suppress the dependence on S for notational convenience. Let $s=|S|$ and

$$t=\frac{\phi^2_\Theta(S)}{32s}.$$

We have

$$|\beta^T(\Sigma-\Theta)\beta|\leq\|\beta\|_1\|(\Sigma-\Theta)\beta\|_\infty\leq t\|\beta\|_1^2,$$

where we applied Hölder twice.

If $\|\beta_N\|\leq 3\|\beta_S\|_1$, then we have

$$\|\beta\|_1\leq 4\|\beta_S\|_1\leq 4\frac{\sqrt{\beta^T\Theta\beta}}{\phi_\Theta/\sqrt{s}}.$$

Thus, we have

$$\beta^T\Theta\beta-\frac{\phi^2_\Theta}{32s}\cdot\frac{16\beta^T\Theta\beta}{\phi^2_\Theta/s}=\frac{1}{2}\beta^T\Theta\beta\leq\beta^T\Sigma\beta. \qquad \square$$

Theorem. Suppose the rows of X are iid and each entry is sub-Gaussian with parameter v. Suppose $s\sqrt{\log p/n}\to 0$ as $n\to\infty$, and $\phi^2_{\Sigma^0,s}$ is bounded away from 0. Then if $\Sigma^0=\mathbb{E}\hat{\Sigma}$, then we have

$$\mathbb{P}\left(\phi^2_{\hat{\Sigma},s}\geq\frac{1}{2}\phi^2_{\Sigma^0,s}\right)\to 1 \text{ as } n\to\infty.$$

Proof. It is enough to show that

$$\mathbb{P}\left(\max_{jk}|\hat{\Sigma}_{jk}-\Sigma^0_{jk}|\leq\frac{\phi^2_{\Sigma^0,s}}{32s}\right)\to 0$$

as $n\to\infty$.

Let $t=\frac{\phi^2_{\Sigma^0,s}}{32s}$. Then

$$\mathbb{P}\left(\max_{j,k}|\hat{\Sigma}_{jk}-\Sigma^0_{jk}|\geq t\right)\leq\sum_{j,k}\mathbb{P}(|\hat{\Sigma}_{jk}-\Sigma^0_{jk}|\geq t).$$

Recall that

$$\hat{\Sigma}_{jk}=\frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik}.$$

So we can apply Bernstein's inequality to bound

$$\mathbb{P}(|\hat{\Sigma}_{jk}-\Sigma^0_{jk})\leq 2\exp\left(-\frac{nt^2}{2(64v^4+4v^2t)}\right),$$

since $\sigma=8v^2$ and $b=4v^2$. So we can bound

$$\mathbb{P}\left(\max_{j,k}|\hat{\Sigma}_{jk}-\Sigma^0_{jk}|\geq t\right)\leq 2p^2\exp\left(-\frac{cn}{s^2}\right)=2\exp\left(-\frac{cn}{s^2}\left(c-\frac{2s^2}{n\log p}\right)\right)\to 0$$

for some constant c. □

Corollary. Suppose the rows of X are iid mean-zero multivariate Gaussian with variance Σ^0. Suppose Σ^n has minimum eigenvalue bounded from below by $c_{min}>0$, and suppose the diagonal entries of Σ^0 are bounded from above. If $s\sqrt{\log p/n}\to 0$, then

$$\mathbb{P}\left(\phi^2_{\hat{\Sigma},s}\geq\frac{1}{2}c_{min}\right)\to 1 \text{ as } n\to\infty.$$

3.6 Computation of Lasso solutions

3.7 Extensions of the Lasso

4 Graphical modelling

4.1 Conditional independence graphs

Proposition. If P has a positive density, then if it satisfies the pairwise Markov property with respect to $\mathcal{G}$, then it also satisfies the global Markov property.

Proposition. Suppose $Z \sim N_p(\mu, \Sigma)$ and Σ is positive definite. Then

$$Z_A \mid Z_B = z_B \sim N_{|A|}(\mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(z_B - \mu_B), \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}).$$

Proof. Of course, we can just compute this directly, maybe using moment generating functions. But for pleasantness, we adopt a different approach. Note that for any M, we have

$$Z_A = MZ_B + (Z_A - MZ_B).$$

We shall pick M such that $Z_A - MZ_B$ is independent of Z_B, i.e. such that

$$0 = \operatorname{cov}(Z_B, Z_A - MZ_B) = \Sigma_{B,A} - \Sigma_{B,B}M^T.$$

So we should take

$$M = (\Sigma_{B,B}^{-1}\Sigma_{B,A})^T = \Sigma_{A,B}\Sigma_{B,B}^{-1}.$$

We already know that $Z_A - MZ_B$ is Gaussian, so to understand it, we only need to know its mean and variance. We have

$$\begin{aligned}
\mathbb{E}[Z_A - MZ_B] &= \mu_A - M\mu_B = \mu_A - \Sigma_{AB}\Sigma_{BB}^{-1}\mu_B\\
\operatorname{var}(Z_A - MZ_B) &= \Sigma_{A,A} - 2\Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A} + \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}\\
&= \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}.
\end{aligned}$$

Then we are done. □

Lemma. Given k, let j' be such that $(Z_{-k})_j = Z_{j'}$. This j' is either j or $j+1$, depending on whether it comes after or before k.

If the jth component of $\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}$ is 0, then $Z_k \amalg Z_{j'} \mid Z_{-kj'}$.

Proof. If the jth component of $\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}$ is 0, then the distribution of $Z_k \mid Z_{-k}$ will not depend on $(Z_{-k})_j = Z_{j'}$ (here j' is either j or $j+1$, depending on where k is). So we know

$$Z_k \mid Z_{-k} \overset{d}{=} Z_k \mid Z_{-kj'}.$$

This is the same as saying $Z_k \amalg Z_{j'} \mid Z_{-kj'}$. □

Lemma. Let $M \in \mathbb{R}^{p\times p}$ be positive definite, and write

$$M = \begin{pmatrix} P & Q \\ Q^T & R \end{pmatrix},$$

where P and Q are square. The *Schur complement* of R is

$$S = P - QR^{-1}Q^T.$$

Note that this has the same size as P. Then

(i) S is positive definite.

(ii)
$$M^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}QR^{-1} \\ -R^{-1}Q^TS^{-1} & R^{-1} + R^{-1}Q^TS^{-1}QR^{-1} \end{pmatrix}.$$

(iii) $\det(M) = \det(S)\det(R)$

4.2 Structural equation modelling

Proposition. If P has a density with respect to a product measure, then (i) and (ii) are equivalent.

Proposition. Let P be the structural equation model with DAG $\mathcal{G}$. Then P obeys the Markov factorization property.

Proof. We assume $\mathcal{G}$ is topologically ordered (i.e. the identity map is a topological ordering). Then we can always write

$$f(z_1, \dots, z_p) = f(z_1)f(z_2 \mid z_1)\cdots z(z_p \mid z_1, z_2, \dots, z_{p-1}).$$

By definition of a topological order, we know $\mathrm{pa}(k) \subseteq \{1, \dots, k-1\}$. Since Z_k is a function of $Z_{\mathrm{pa}(k)}$ and independent noise ε_k. So we know

$$Z_k \amalg Z_{\{1,\dots,p\}\setminus\{k\cup\mathrm{pa}(k)\}} \mid Z_{\mathrm{pa}(k)}.$$

Thus,
$$f(z_k \mid z_1, \dots, z_{k-1}) = f(z_k \mid z_{\mathrm{pa}(k)}).$$ □

4.3 The PC algorithm

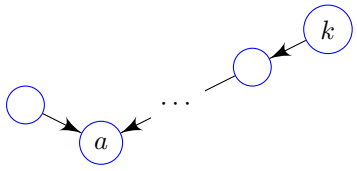
Proposition. Two DAGs are Markov equivalent iff they have the same skeleton and same set of v-structure.

Proposition. If nodes j and k are adjacent in a DAG $\mathcal{G}$, then no set can d-separate them.

If they are not adjacent, and π is a topological order for $\mathcal{G}$ with $\pi(j) < \pi(k)$, then they are d-separated by $\mathrm{pa}(k)$.

Proof. Only the last part requires proof. Consider a path $j = j_1, \dots, j_m = k$. Start reading the path from k and go backwards. If it starts as $j_{m-1} \to k$, then j_{m-1} is a parent of k and blocks the path. Otherwise, it looks like $k \to j_{m-1}$.

We keep going down along the path until we first see something of the form



Thus must exist, since j is not a descendant of k by topological ordering. So it suffices to show that a does not have a descendant in $\mathrm{pa}(k)$, but if it did, then this would form a closed loop. □

Proposition. Suppose we have $j - \ell - k$ in the skeleton of a DAG.

(i) If $j \to \ell \leftarrow k$, then no S that d-separates j can have $\ell \in S$.

(ii) If there exists S that d-separates j and k and $\ell \notin S$, then $j \to \ell \leftarrow k$. $\square$

5 High-dimensional inference

5.1 Multiple testing

Theorem. When using the Bonferroni correction, we have

$$\text{FWER} \leq \mathbb{E}(N) \leq \frac{m_0\alpha}{m} \leq \alpha.$$

Proof. The first inequality is Markov's inequality, and the last is obvious. The second follows from

$$\mathbb{E}(N) = \mathbb{E}\left(\sum_{i\in I_0} \mathbf{1}_{p_i\leq \alpha/m}\right) = \sum_{i\in I_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{m}\right) \leq \frac{m_0\alpha}{m}. \qquad \square$$

Theorem. Closed testing makes no false rejections with probability $\geq 1-\alpha$. In particular, FWER $\leq \alpha$.

Proof. In order for there to be a false rejection, we must have falsely rejected H_{I_0} with the local test, which has probability at most α. $\square$

Theorem. Suppose that for each $i \in I_0$, p_i is independent of $\{p_j : j \neq i\}$. Then using the Benjamini–Hochberg procedure, the false discovery rate

$$FDR = \mathbb{E}\frac{N}{\max(R,1)} \leq \frac{\alpha M_0}{M} \leq \alpha.$$

Proof. The false discovery rate is

$$\begin{aligned}
\mathbb{E}\frac{N}{\max(R,1)} &= \sum_{r=1}^{M} \mathbb{E}\frac{N}{r}\mathbf{1}_{R=r} \\
&= \sum_{r=1}^{m} \frac{1}{r}\mathbb{E}\sum_{i\in I_0} \mathbf{1}_{p_i\leq \alpha r/M}\mathbf{1}_{R=r} \\
&= \sum_{i\in I_0}\sum_{r=1}^{M} \frac{1}{r}\mathbb{P}\left(p_i \leq \frac{\alpha r}{m}, R = r\right).
\end{aligned}$$

The brilliant idea is, for each $i \in I_0$, let R_i be the number of rejections when applying a modified Benjamini–Hochberg procedure to $p^{\setminus i} = \{p_1, \dots, p_M\} \setminus \{p_i\}$ with cutoff

$$\hat{k}_i = \max\left\{j : p^{\setminus i}_{(j)} \leq \frac{\alpha(j+1)}{m}\right\}$$

We observe that for $i \in I_0$ and $r \geq 1$, we have

$$\begin{aligned}
\left\{p_i \leq \frac{\alpha r}{m},\, R = r\right\} &= \left\{p_i \leq \frac{ar}{m},\, p_{(r)} \leq \frac{\alpha r}{m},\, p_{(s)} > \frac{\alpha s}{m} \text{ for all } s \geq r\right\} \\
&= \left\{p_i \leq \frac{\alpha r}{m},\, p^{\setminus i}_{(r-1)} \leq \frac{\alpha r}{m},\, p^{\setminus i}_{(s-1)} > \frac{\alpha s}{m} \text{ for all } s > r\right\} \\
&= \left\{p_i \leq \frac{\alpha r}{m},\, R_i = r-1\right\}.
\end{aligned}$$

The key point is that $R_i = r - 1$ depends only on the other p-values. So the FDR is equal to

$$\begin{aligned}\text{FDR} &= \sum_{i\in I_0}\sum_{r=1}^{M}\frac{1}{r}\mathbb{P}\left(p_i \le \frac{\alpha r}{M}, R_i = r-1\right)\\ &= \sum_{i\in I_0}\sum_{r=1}^{M}\frac{1}{r}\mathbb{P}\left(p_i \le \frac{\alpha r}{m}\right)\mathbb{P}(R_i = r-1)\end{aligned}$$

Using that $\mathbb{P}(p_i \le \frac{\alpha r}{m}) \le \frac{\alpha r}{m}$ by definition, this is

$$\begin{aligned}&\le \frac{\alpha}{M}\sum_{i\in I_0}\sum_{r=1}^{m}\mathbb{P}(R_i = r-1)\\ &= \frac{\alpha}{M}\sum_{i\in I_0}\mathbb{P}(R_i \in \{0,\dots,m-1\})\\ &= \frac{\alpha M_0}{M}.\end{aligned}$$

□

5.2 Inference in high-dimensional regression

Theorem. Suppose the maximum eigenvalue of Σ is always at least $c_{min} > 0$ and $\max_j \Sigma_{jj} \le 1$. Suppose further that $s_{max}\sqrt{\log(p)/n} \to 0$. Then there exists constants A_1, A_2 such that setting $\lambda = \lambda_j = A_1\sqrt{\log(p)/n}$, we have

$$\begin{aligned}\sqrt{n}(\hat{b} - \beta^0) &= W + \Delta\\ W \mid X &\sim N_p(0, \sigma^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^T),\end{aligned}$$

and as $n, p \to \infty$,

$$\mathbb{P}\left(\|\Delta\|_\infty > A_2 s\frac{\log(p)}{\sqrt{n}}\right) \to 0.$$

Proof. Consider the sequence of events Λ_n defined by the following properties:

- $\phi_{\hat{\Sigma},s} \ge c_{min}/2$ and $\phi^2_{\hat{\Sigma}_{-j,-j},s_j} \ge c_{min}/2$ for all j
- $\frac{2}{n}\|X^T\Sigma\|_\infty \le \lambda$ and $\frac{2}{n}\|X^T_{-j}\varepsilon^{(j)}\|_\infty \le \lambda$.
- $\frac{1}{n}\Sigma^{(j)}\|_2^2 \ge (\Omega_{jj})^{-1}(1 - 4\sqrt{(\log p)/n})$

Question 13 on example sheet 4 shows that $\mathbb{P}(\Lambda_n) \to 1$ for A_1 sufficiently large. So we will work on the event Λ_n.

By our results on the Lasso, we know

$$\|\beta^0 - \hat{\beta}\|_1 \le c_1 s\sqrt{\log p/n}.$$

for some constant c_1. We now seek a lower bound for $\hat{\tau}_j^2$. Consider linear models

$$X_j = X_{-j}\gamma^{(j)} + \varepsilon^{(j)},$$

where the sparsity of $\gamma^{(j)}$ is s_j, and $\varepsilon_i^{(j)}|X_{-j} \sim N(0, \Omega_{jj}^{-1})$. Note that

$$\Omega_{jj}^{-1} = \text{var}(X_{ij} \mid X_{i,-j}) \le \text{var}(X_{ij}) = \Sigma_{ij} \le A.$$

Also, the maximum eigenvalue of Ω is at most c_{min}^{-1}. So $\Omega_{jj} \leq c_{min}^{-1}$. So $\Omega_{jj}^{-1} \geq c_{min}$. So by Lasso theory, we know

$$\|\gamma^{(j)} - \hat{\gamma}^{(j)}\|_1 \leq c_2 s_j \sqrt{\frac{\log p}{n}}$$

for some constant c_2. Then we have

$$\begin{aligned}
\hat{\tau}_j^2 &= \frac{1}{n}\|X_j - X_{-j}\hat{\gamma}^{(j)}\|_2^2 + \lambda\|\hat{\gamma}^{(j)}\|_1 \\
&\geq \frac{1}{n}\|\varepsilon^{(j)} + X_{-j}(\gamma^{(j)} - \hat{\gamma}^{(j)})\|_2^2 \\
&\geq \frac{1}{n}\|\varepsilon^{(j)}\|_2^2 - \frac{2}{n}\|X_{-j}^T\varepsilon^{(j)}\|_\infty\|\gamma^{(j)} - \hat{\gamma}^{(j)}\|_1 \\
&\geq \Omega_{jj}^{-1}\left(1 - 4\sqrt{\frac{\log p}{n}}\right) - c_2 s_j\sqrt{\frac{\log p}{n}} + A_1\sqrt{\frac{\log p}{n}}
\end{aligned}$$

In the limit, this tends to Ω_{jj}^{-1}. So for large n, this is $\geq \frac{1}{2}\Omega_{jj}^{-1} \geq \frac{1}{2}c_{min}$.

Thus, we have

$$\|\Delta\|_\infty \leq 2\lambda\sqrt{n}c_1 s\sqrt{\frac{\log p}{n}}c_{min}^{-1} = A_2 s\frac{\log p}{\sqrt{n}}. \qquad \square$$