Modern Statistical Methods Example Sheet 1 Solutions

Autumn Mapes

Fall 2025

1 Chapter 0: Preliminaries

1.1 Some Important Identities

Def: Law of Total Expectation

AKA the tower property

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$$

Card ID: 1769639660396

Def: Chain Rule of Probability

$$\begin{aligned}\mathbb{P}(X,Y) &= \mathbb{P}(Y)\mathbb{P}(X|Y)\\ &= \mathbb{P}(X)\mathbb{P}(Y|X)\end{aligned}$$

Card ID: 1769639660397

Def: Slutsky's Theorem

Let X_n, Y_n be sequences of random elements. If both
$X_n \xrightarrow{d} X$
$Y_n \xrightarrow{p} c$
for some constant c, then

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} Xc$
3. $X_n / Y_n \xrightarrow{d} X/c$ given $c \neq 0$.

Card ID: 1769639660400

2 Chapter 1: Ridge Regression and the Kernel Trick

2.1 Ridge Regression

Vector generalization of the mean squared error (MSE) for an estimator $\hat{\theta}$ of θ, and what it simplifies down to

---

$$\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = \text{Var}(\hat{\theta}) + Bias(\hat{\theta})Bias(\hat{\theta})^T = \text{Var}(\hat{\theta}) + \mathbb{E}[\hat{\theta} - \theta]\mathbb{E}[\hat{\theta} - \theta]^T$$

Card ID: 1778792585818

Def: Equation optimized for Ridge Regression

$$(\hat{\mu}_\lambda^R, \hat{\beta}_\lambda^R) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{argmin} \; (||Y - \mu 1 - X\beta||_2^2 + \lambda||\beta||_2^2)$$

Card ID: 1778792585820

Def: Solution for $\hat{\beta}_\lambda^R$ in Ridge Regression

$$(X^T X + \lambda I)^{-1} X^T Y$$

Card ID: 1778792585821

What do we know about the differences between OLS solutions β^{OLS} and ridge regression solutions β^R?

As $\lambda \to 0$, the difference in MSEs between β^{OLS} and β^R_λ is positive definite:

$$\mathbb{E}[(\beta^{OLS} - \beta_0)(\beta^{OLS} - \beta_0)^T] - \mathbb{E}[(\beta^R - \beta_0)(\beta^R - \beta_0)^T] \succeq 0$$

Show this by first computing the bias and variance of $\hat{\beta}^R_\lambda$.

Card ID: 1778792585823

In MSM, what does the notation X_j mean?

The jth column of X

Card ID: 1778792585825

What does SVD tell about when ridge regression works best?

λ shrinks Y most in the small principal components of X. So if most of the signal is in the large principal components, we won't lose much.

Card ID: 1778792585826

2.2 The Kernel Trick

Show how the kernel trick leads to an alternate form of $\hat{\beta}^R_\lambda$

Note first that for $X \in \mathbb{R}^{n \times p}$,

$$(X^T X + \lambda I_n) X^T = X^T (X X^T + \lambda I_p)$$

Multiply by inverses on each side and then multiplying by Y on the right yields

$$(X^T X + \lambda I_n)^{-1} X^T Y = X^T (X X^T + \lambda I_p)^{-1} Y = \hat{\beta}^R_\lambda$$

Note that the fitted values $X\hat{\beta}^R_\lambda$ only depend on XX^{T} then!

Card ID: 1778792585828

Kernel Form of Ridge Regression

$$X\hat{\beta}_\lambda^R = K(K + \lambda I)^{-1}Y$$

where $K = XX^{\mathrm{T}}$

Card ID: 1778792585830

What three properties must inner products have?

1. Conjugate Symmetry: $\langle x, y\rangle = \overline{\langle y, x\rangle}$
2. Linearity in the first argument: $\langle ax + by, z\rangle = a\langle x, y\rangle + b\langle x, y\rangle$
3. Positive Definiteness: $\langle x, x\rangle \geq 0$, equality iff $x = 0$

for $x, y \in \mathcal{H}$, $a, b \in \mathbb{K}$
Note that 1 and 2 together makes the inner product a sesquilinear form.

Card ID: 1778792585831

Feature Map

A function $\phi : X \to H$ from X to some Hilbert space H.

Note that this gives us a similarity measure $k : X \times X \to \mathbb{R}$:

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

Card ID: 1778792585832

Def: Positive Definite Kernel

A symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that the matrix K with entries

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

is positive semi-definite.

Card ID: 1778792585834

Def: Cauchy-Schwarz for Kernels

$$k(x, x')^2 \leq k(x, x)k(x', x')$$

Card ID: 1778792585836

Def: Sum form of positive-semi-definiteness

A matrix X is positive semi-definite iff

$$\sum_{i,j} a_i X_{ij} a_j \geq 0$$

Card ID: 1778890950638

Def: Kernel from a Feature Map

Given some feature map $\phi : \mathcal{X} \to \mathcal{H}$, the following is a kernel:

$$k(x, y) = \langle \phi(x), \phi(y) \rangle$$

This is because

$$\sum_{i,j} a_i k(x_i, x_j) a_j = \sum_{i,j} a_i \langle \phi(x_i), \phi(x_j) \rangle a_j = \langle \sum_i a_i \phi(x_i), \sum_i a_i \phi(x_i) \rangle \geq 0$$

Card ID: 1778890950640

What does a matrix being positive definite tell us about its eigenvalues?

The eigenvalues are positive (and real).

Card ID: 1778792585838

What is the definition of a matrix being positive-definite?

A IS SYMMETRIC, and for any nonzero real column vector x,
$x^T A x > 0$

Note this implies

$$\sum_{i,j} x_i A_{ij} x_j > 0$$

Semidefinite just means it can also be zero.

Card ID: 1778792585839

If some matrix A has full column rank, what does that tell us about $A^T A$?

$A^T A$ is positive definite

Card ID: 1778792585841

What are the three kernel closure properties?

---

1. Closure over addition and nonnegative scalars: $a_1 k_1 + a_2 k_2$
2. Closure over sequences: $\lim_{k\to\infty} k_m$
3. Closure over products: $k_1 k_2$

Card ID: 1778792585842

Def: Linear Kernel

$k(x, x') := x^{\mathrm{T}} x'$

Card ID: 1778792585844

Def: Polynomial Kernel

$k(x, x') := (1 + x^{\mathrm{T}} x')^d$

Card ID: 1778792585845

Def: Gaussian Kernel

$k(x, x') := \exp\left(-\frac{||x-x'||_2^2}{2h^2}\right)$ $h > 0$ is called the bandwidth.

Note that this must be infinite-dimensional because we have a term $e^{xx'}$, which Taylor expands to an infinite series of polynomial terms.

Card ID: 1778792585847

Def: First-Order Sobolev Kernel

$$k(x, x') = \min(x, x')$$

This must be a kernel because it can be represented as

$$k(x, x') = \int_0^1 \mathbb{1}_{[0,x]}(u)\mathbb{1}_{[0,x']}(u) = \langle \mathbb{1}_{[0,x]}, \mathbb{1}_{[0,x']} \rangle_{L_2([0,1])}$$

Card ID: 1778792585848

Def: Second-Order Sobolev Kernel

$$k(x, x') = \int_0^{\min(x,x')} (x - u)(x' - u)du$$

Card ID: 1778792585850

Def: Jaccard Similarity

Similarity based on what features are shared among $1...p$. $h(x, x')$ is defined as 1 when $x \cup x' = \emptyset$, and otherwise

$$h(x, x') = \frac{|x \cap x'|}{|x \cup x'|}$$

where $|.|$ is the set size.

Card ID: 1778792585851

2.3 Reproducing Kernel Hilbert Spaces

Def: The Spectral Theorem

If X is symmetric, we have a decomposition

$$X = P^{\mathrm{T}} D P$$

where D is diagonal and P is orthogonal. The rows of P are the eigenvectors of X.

Furthermore, if X is positive definite, every diagonal term in D is positive.

Card ID: 1778792585852

Prop

For every kernel k on FINITE $\mathcal{X}$, there exists some Hilbert space $\mathcal{H}$ and feature map $\phi : \mathcal{X} \to \mathcal{H}$ for which

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

Because $\mathcal{X}$ is finite and k is a kernel, we have a positive definite $K \in \mathbb{R}^{n \times n}$ that contains all information about the space. Because K is positive definite, we have the decomposition $K = P^{\mathrm{T}} D P$ with orthogonal P and positive diagonal D. Take $\mathcal{H}$ to be $\mathbb{R}^n$ and the feature map $\phi : x_i \mapsto (D^{1/2} P_i)$

Card ID: 1778792585854

Prop

For every kernel k in a general $\mathcal{X}$, there exists some Hilbert space $\mathcal{H}$ and feature map $\phi : \mathcal{X} \to \mathcal{H}$ for which

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

First specify the $\mathcal{H}$ and $\phi(x)$ used.

Take $\mathcal{H}$ to be the linear span of the functions

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

for some $n \in \mathbb{N}$, $x_i \in \mathcal{X}$, and $a_i \in \mathbb{R}$ with the inner product defined as

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

This is clearly well-defined since there's no dependence on the expansion of f or g. So our feature map is

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

Card ID: 1778792585856

Prop

Show that $\mathcal{H}$ defined as the space of functions of the form

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

with inner product

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

and

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

is a valid kernel and inner product.

First show that we indeed have

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

Clearly symmetric and linear, but is it positive definite?
Note first that $\langle f, g \rangle \geq 0$ by positive definiteness of the kernel. So then for functions $f_1 ... f_n$, coefficients $\gamma_1 ... \gamma_n$, and working from the PSD requirement, we have:

$$\sum_{i,j} \gamma_i \langle f_i, f_j \rangle \gamma_j = \left\langle \sum_i f_i \gamma_i, \sum_j f_j \gamma_j \right\rangle \geq 0$$

Thus $\langle \cdot, \cdot \rangle$ is a kernel. The last part to show is that $\langle f, f \rangle = 0 \iff f = 0$. The backwards direction is obvious, but for the forwards direction, first note that

$$\langle k(\cdot, x), f \rangle = \sum_i a_i k(x_i, x) = f(x)$$

So, we have:

$$\langle f, f \rangle = 0 \implies f(x) = \langle k(\cdot, x), f \rangle \leq \langle k(\cdot, x), k(\cdot, x) \rangle \langle f, f \rangle = 0$$

where the $\leq$ step is given by Cauchy-Schwarz for kernels.

Card ID: 1778792585858

Def: Reproducing Kernel Hilbert Space (RKHS)

A Hilbert space of functions $f : X \to \mathbb{R}$ such that, for all $x \in X$,

$$f(x) = \langle f, K_x \rangle$$

Note: this condition is equivalent to saying that the point-evaluation functional

$$L_x = H \to \mathbb{R},\ L_x(f) = f(x)$$

is continuous for every $x \in X$.
So in other words $L_x \in H^*$
This is equivalent to the condition above by the Riesz Representation Theorem

Card ID: 1768311461712

Prop

The reproducing kernel of an RKHS is well-defined.

Given any two reproducing kernels $k_x, h_x \in X \times X \to \mathbb{R}$, show that

$$||k_x - h_x||_H^2 = 0$$

by representing as an inner product and simplifying:

$$\begin{aligned} ||k_x - h_x||_H^2 &= \langle k_x - h_x, k_x - h_x \rangle \\ &= \langle k_x, k_x - h_x \rangle - \langle h_x, k_x - h_x \rangle \\ &= (k_x - h_x)(x) - (k_x - h_x)(x) \\ &\text{by the reproducing property} \\ &= 0 \end{aligned}$$

Card ID: 1768435499085

Def: Reproducing kernel (of an RKHS)

By definition, fixing some $x \in X$, there exists some K_x such that for any $f \in H$, $f(x) = \langle f, K_x \rangle$.
If we plug in some y to K_x, we then get $K_x(y) = \langle K_x, K_y \rangle$
So our reproducing kernel is the function where we take some x, y pair and perform that process:

$$K(x, y) = \langle K_x, K_y \rangle$$

$K : X \times X \to \mathbb{R}$

Card ID: 1768311941362

2.4 The Representer Theorem

State the Representer Theorem

Given some RKHS $\mathcal{H}$ with representing kernel K and arbitrary loss function $c : \mathbb{R}^n \times X^n \times \mathbb{R}^n \to \mathbb{R}$ and strictly increasing $J : \mathbb{R} \to \mathbb{R}$, the minimizing $\hat{f} \in \mathcal{H} : X \to \mathbb{R}$ of

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

is given by

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, x_i)$$

with $\hat{\alpha}$ minimizing:

$$Q_2(\alpha) = c(y, x_1, ..., x_n, K\alpha) + J(\alpha^T K \alpha)$$

Card ID: 1778792585861

Prop

Show the representer theorem:
Given some RKHS $\mathcal{H}$ with representing kernel K and arbitrary loss function $c : \mathbb{R}^n \times X^n \times \mathbb{R}^n \to \mathbb{R}$ and strictly increasing $J : \mathbb{R} \to \mathbb{R}$, the minimizing $\hat{f} \in \mathcal{H} : X \to \mathbb{R}$ of

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

is given by

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, x_i)$$

with $\hat{\alpha}$ minimizing:

$$Q_2(\alpha) = c(y, x_1, ..., x_n, K\alpha) + J(\alpha^T K \alpha)$$

Let U be the linear span of $k(\cdot, x_i), i = 1...n$. Linear subspaces of Hilbert spaces are closed; and closed subspaces of Hilbert spaces mean that any function $f = u + v$: $u \in U$ and $v \in U^{\perp}$.
Show first that $f(x_i) = u(x_i)$ by means of the definition of RKHS, and that $J(||f||^2_{\mathcal{H}}) \geq J(||u||^2_{\mathcal{H}})$. Reason then that Q_1 is minimized when $v = 0$.
Show then that we can express $||u||^2_{\mathcal{H}}$ as $\alpha^T K \alpha$ and $f(x_i)$ as $K\alpha$; thus $Q_1(f) = Q_2(\alpha)$.

Card ID: 1778792585863

2.5 Kernel Ridge Regression

Functional Generalization of Ridge Regression

Find the $\hat{f}_\lambda$ minimizing

$$\sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda ||f||^2_{\mathcal{H}}$$

Card ID: 1778792585864

How can we simplify

$$\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda||f||_{\mathcal{H}}^2$$

with the Representer Theorem?

Using $c(...) = \sum_{i=1}^{n}(Y_i - f(x_i))^2$ and $J(||f||_{\mathcal{H}}^2) = \lambda||f||_{\mathcal{H}}^2$, we have that the above is equivalent to minimizing

$$||Y - K\alpha||_2^2 + \lambda\alpha^{\mathrm{T}}K\alpha$$

Note that this gives us a way to predict an unknown point:

$$\hat{f}(x) = \sum_{i=1}^{n}\hat{\alpha}_i k(x, x_i)$$

Card ID: 1778792585865

What do we know about orthogonal matrices in the L2 norm?

Because an orthogonal matrix never scales anything, we have:

$$||UX||_2^2 = ||X|_2^2 = ||XU||_2^2$$

Card ID: 1778792585867

Trick to pull orthogonal matrices out of an inverse like $(UDU^{\mathrm{T}} + \lambda I)^{-1}$

Note that $I = U^{\mathrm{T}}U = UU^{\mathrm{T}}$, so we have:

$$\begin{aligned}(UDU^{\mathrm{T}} + \lambda I)^{-1} &= (UDU^{\mathrm{T}} + U\lambda IU^{\mathrm{T}})^{-1} \\ &= (U(D + \lambda I)U^{\mathrm{T}})^{-1} \\ &= U^{-T}(D + \lambda I)^{-1}U^{-1} \\ &= U(D + \lambda I)U^{\mathrm{T}}\end{aligned}$$

Card ID: 1778792585868

AM-GM

The arithmetic mean is always at least as large as the geometric mean: for all $a, b \in \mathbb{R}$,

$$\frac{a+b}{2} \geq \sqrt{ab}$$

Card ID: 1778792585870

How can we bound a term like $\frac{ab}{(a+b)^2}$?

Use AM-GM:

$$\frac{a+b}{2} \geq \sqrt{ab}$$

$$a+b \geq 2\sqrt{ab}$$

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

$$\frac{ab}{(a+b)^2} \leq \frac{ab}{4ab} = \frac{1}{4}$$

Card ID: 1778792585872

Thm: Prove that the mean squared prediction error (MSPE) may be bounded above in the following way:

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

Proof sketch:
By the representer theorem, we have that

$$[\hat{f}_\lambda^R(x_i)] = K(K+\lambda I)^{-1}Y$$

and

$$[\hat{f}_\lambda^0(x_i)] = K\alpha$$

Use the eigendecomposition $K = UDU^{\mathrm{T}}$, U orthogonal.
Let $\Theta = DU^{\mathrm{T}}\alpha$, note that $Y = K\alpha+\varepsilon$. Split into deterministic and stochastic parts (and swap them, stochastic is the first term in the final representation)
In the deterministic term, use the D^{-1} trick to introduce $\alpha^{\mathrm{T}}K\alpha = ||f^0||_{\mathcal{H}}^2$; get the $1/4$ from bounded the fraction with AM-GM.
In the stochastic bit, use the trace trick to get the ϵs together, where you can bring the expectation in to get σ^2.
The final equality falls out of AM-GM again.

Card ID: 1778792585873

Def: Eigenfunction

Given a random variable X taking values in $\mathcal{X}$, we say a nonzero function $e \in \mathcal{H}$ is an eigenfunction with eigenvalue $\mu \in \mathbb{R}$ if

$$\mu e(x) = \mathbb{E}[k(x,X)e(X)]$$

in other words, the $\mathbb{E}[k(x,X)...]$ part is a linear operator acting upon e:

$$(Te)(x) = \int k(x,x')e(x')dP(x')$$

Card ID: 1778792585875

Def: Mercer's Theorem

Given a random variable X taking values in $\mathcal{X}$ with a nonzero $e \in \mathcal{H}$ eigenfunction with eigenvalue $\mu \in \mathbb{R}$ (so that $\mu e(x) = \mathbb{E}[k(x, X)e(X)]$ is fulfilled), and under some mild regularity conditions (like $\mathbb{E}(k(X, X)) < \infty$), then three things must hold true:

1. The set of positive eigenvalues is at most countable.

2. The subspace spanned by each eigenfunction has a finite dimension (known as the multiplicity of the eigenvalue)

3. The eigenvalues are orthonormal in the sense that

$$\mathbb{E}[e_j(X)e_k(X)] = \mathbb{1}_{j=k}$$

See the decomposition of kernels.

Card ID: 1778792585877

Mercer's Theorem decomposition for kernels

$$k(x, y) = \sum_{j \in J} \mu_j e_j(x) e_j(y)$$

Card ID: 1779648618272

TODO TECHNICAL MERCER'S THEOREM STUFF

2.6 Large-Scale Kernel Machines

Def: Bochner's Theorem

Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant iff there exists some $c > 0$ and distribution F on $\mathbb{R}^p$ such that when $W \sim F$,

$$k(x, x') = c\mathbb{E}[e^{i(x-x')^{\mathrm{T}}W}] = c\mathbb{E}[\cos((x - x')^{\mathrm{T}}W)]$$

Card ID: 1778792585879

TODO BOCHNER'S THEOREM EXAMPLE

3 Chapter 2: The Lasso

3.1 The Lasso

Def: Mean-Squared Prediction Error (MSPE)

$$\frac{1}{n}\mathbb{E}[||X\beta^0 - X\hat{\beta}||_2^2]$$

Measures the average difference between true values and predicted values.

Card ID: 1778792585881

Why do we expect that Lasso regression could be useful?

Let S be the set $\{k : \beta_k^0 \neq 0\}$ (the set of parameters that are nonzero). Oftentimes this set is smaller than p, the total number of parameters. When this is the case, OLS could stand to be a lot better: consider the MSPE:

$$\frac{1}{n}\mathbb{E}[||X\beta^0 - X\hat{\beta}||_2^2] = ... = \frac{\sigma^2 p}{n}$$

If we just dropped each of the ps that aren't in S, we could significantly reduce the error.

Card ID: 1778792585882

Function minimized in Lasso regression

Find the $(\hat{\mu}^L, \hat{\beta}^L)$ minimizing

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

We generally assume that X is already mean-centered, giving us the loss we actually seek to minimize:

$$Q_\lambda(\beta) := \frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$$

Card ID: 1778792585884

Given the minimizer $\hat{\beta}_\lambda^L$ in lasso regression, how can we express this as a conditional optimization problem?

$$||Y - X\beta||_2^2 \text{ subject to } ||\beta||_1 \leq ||\hat{\beta}_\lambda^L||_1$$

Card ID: 1778792585886

Def: Holder's Inequality

Let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then we have, for any functions f, g,

$$||fg||_1 \leq ||f||_p ||g||_q$$

Card ID: 1778792585888

Def: Basic Inequality for the Lasso

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

where $\hat{\beta}$ is the true optimized solution of the Lasso loss equation.

Card ID: 1779140587966

Thm: Prediction Error of the Lasso (slow rate):
Let $\hat{\beta}$ be any Lasso solution when

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

with probability at least $1 - 2p^{-(A^2/2-1)}$. Then:

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

Proof sketch:
Start from the fact that we know β^L can do no better than the true mean β^0:

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

Substitute in $Y = X\beta^0 + \varepsilon$, match the left side.
We can put an implicit norm around $\varepsilon^{\mathrm{T}}X(\hat{\beta} - \beta^0)$, use Holder's to get the one norm we need paired with an infinity norm around $X^{\mathrm{T}}\varepsilon$
In order to make everything work out nicely, we need the assumption that $\frac{1}{n}||X^{\mathrm{T}}\varepsilon||_\infty \leq \lambda$. Call this event Ω, we assume that it is true with probability at least $1 - 2p^{-(A^2/2-1)}$.
Then just use the triangle inequality and multiply both sides by two.
Card ID: 1778792585889

3.2 Concentration Inequalities I

State Markov's Inequality

If X is a nonnegative random variable and $a > 0$, then we have:
$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$
Card ID: 1778792585892

State Chebyshev's Inequality

$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$
Note that this is useless when $k \leq 1$, since all probabilities are ≤ 1

Equivalently we have
$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$
Card ID: 1778792585894

Def: Subgaussian RV

We say a random variable X is subgaussian with parameter σ if:

$$\mathbb{E}[e^{t(X-\mathbb{E}[X])}] \leq e^{t^2\sigma^2/2}$$

Card ID: 1778792585896

Def: Subgaussian Tail Bound

Given X subgaussian with parameter σ, we have:

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq e^{-t^2/(2\sigma^2)}$$

Card ID: 1778792585898

Def: Hoeffding's Lemma

If X takes values in $[a, b]$, then X is subgaussian with parameter $(b-a)/2$.

Card ID: 1778792585900

Def: What must be true about sequences of independent subgaussian RVs?

If independent subgaussian RVs $X_1 ... X_n$ have parameters σ_i respectively, then the RV $\gamma^{\mathrm{T}} X$ is subgaussian with parameters $\left(\sum_i \gamma_i^2 \sigma_i^2\right)^{1/2}$

Card ID: 1778792585902

Thm: Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero and sub-Gaussian with common parameter σ. Note that this includes $\varepsilon \sim N_n(0, \sigma^2 I)$. Let $\lambda = A\sigma\sqrt{\log(p)/n}$. Then

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

Proof sketch:
Start from

$$... \leq \sum_{i=1}^{p} \mathbb{P}(\frac{1}{n}|\varepsilon^T X_i| > \lambda)$$

Split into two events, one where $\varepsilon^T X_i$ is positive, and one where it's negative. Both are subgaussian with parameter $\sigma/\sqrt{n}$, so use the upper bound. Then just plug in λ.

Card ID: 1778792585904

3.3 Convex Analysis

Def: Set Convexity

A set C is convex if, for all $x, y \in C$, $t \in [0, 1]$,

$$tx + (1 - t)y \in C$$

Card ID: 1778890950664

State the definition of function convexity.

$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y)$
for all $x, y \in C$
given convex space C, $f : C \to \mathbb{R}^n$
Card ID: 1778792585906

Write the Lagrangian function as used in the Lagrangian Method to minimize a function $f(x)$ subject to $g(x) = 0$
($f : C \to \mathbb{R}$ ($C \subseteq \mathbb{R}^d$ convex), $g : \mathbb{R}^d \to \mathbb{R}^d$)

$L(x, \theta) = f(x) + \theta^T g(x)$
Find minima in x, θ

We're looking for points where the tangents to the contours of f and g are parallel (or where f alone is minimized and $\theta = 0$)
Card ID: 1778792585908

Given a convex function $f : C \to \mathbb{R}$, when is a vector $v \in \mathbb{R}^d$ a subgradient of f at x_0?

$\forall x \in C$, $f(x) - f(x_0) \geq v^T(x - x_0)$

So in one dimension we have:
$\forall x \in C$, $f(x) - f(x_0) \geq c(x - x_0)$
given $c \in \mathbb{R}^n$
any c such that our function stays above the line drawn out from x with slope c
Card ID: 1778792585909

What is a subdifferential of a function f at x, and how is it denoted?

The set of all subgradients of f at x.
It's denoted $\partial f(x)$
Card ID: 1778792585911

For the interior of a convex region, if f is differentiable, what are the subdifferentials of f?

Just the gradient of f at any point x:
$\partial f(x) = \{\nabla f(x)\}$
Card ID: 1778792585912

What are the simplifying properties of subdifferentials?

---

1. Subdifferentials scale how you'd expect:
$\partial(\alpha f)(x) = \{\alpha v,\, v \in \partial f(x)\}$
2. When adding functions, we consider all permutations:
$\partial(f_1 + f_2)(x) = \{v_1 + v_2,\, v_1 \in \partial f_1(x),\, v_2 \in \partial f_2(x)\}$
3. Subdifferentials ignore constant adds and respect linearity:
$h(x) = f(Ax + b) \implies \partial h(x) = \{A^{\mathrm{T}} g : g \in \partial f(Ax + b)\}$
Card ID: 1778792585914

For a convex function f, how must any minimizer relate to subdifferentials?

We have

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

(both are equivalent to saying $f(y) \leq f(x^*),\, \forall y$)
Card ID: 1778792585916

What is the subdifferential of the L1 norm $\partial ||x||_1$?

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

where $A \in \{j : x_j \neq 0\}$
Card ID: 1778792585918

For $x \in \mathbb{R}^d$, state an equivalent form of $\partial||x||_1$.

$$\partial x||x||_1 = \{v \in \mathbb{R}^d, \, ||v||_\infty \leq 1 \text{ and } v_A = \text{sgn}(x_A)\}$$

where $A = \{j : \, x_j \neq 0\}$

In other words, v where the magnitude of all components of v do not exceed 1, v is positive in the components where x is positive and negative where the components of x is negative.

Card ID: 1778792585920

3.4 KKT Conditions for the Lasso

State the KKT conditions for the Lasso

$$\hat{v} = \frac{1}{\lambda n} X^T (Y - X\beta_\lambda^L)$$

where $\hat{v}$ satisfies $||\hat{v}||_\infty \leq 1$ and $\hat{v}_{\hat{S}} = \text{sgn}(\hat{\beta}^L_{\lambda,\,\hat{S}})$ where $\hat{S} = \{j : \, \beta^L_{\lambda,\,j} \neq 0\}$

Card ID: 1778792585922

What do we know about the fitted values of Lasso minimizers?

They are unique
(though the minimizing β_λ^L may not be unique)

Card ID: 1778792585924

Thm: **For $\lambda \geq 0$, any two lasso regression solutions $\hat{\beta}_\lambda^{(1)}$ and $\hat{\beta}_\lambda^{(2)}$ must have the same fitted values: $X\hat{\beta}_\lambda^{(1)} = X\hat{\beta}_\lambda^{(2)}$**

First note that, by the convexity of $||\cdot||_2^2$ with $t = \frac{1}{2}$,

$$||Y - \frac{1}{2}X\beta_\lambda^{(1)} - \frac{1}{2}X\beta_\lambda^{(2)}||_2^2 \leq ||Y - \frac{1}{2}X\beta_\lambda^{(1)}||_2^2 + ||Y - \frac{1}{2}X\beta_\lambda^{(2)}||_2^2$$

and by strict convexity, equality is achieved iff $X\beta^{(1)} = X\beta^{(2)}$.
Show that

$$c^* \leq Q_\lambda(\frac{1}{2}X\beta^{(1)} + \frac{1}{2}X\beta^{(2)}) \leq \frac{1}{2}Q_\lambda(X\beta^{(1)}) + \frac{1}{2}Q_\lambda(X\beta^{(2)}) = c^*$$

applying convexity of the L1 and L2 norms, and so equality is achieved.
Card ID: 1778792585926

TODO EQUICORRELATION

3.5 Variable Selection

What do the sets S and N denote in the context of Lasso regression?

$S := \{k : \beta_k^0 \neq 0\}$ is the set of nonzero terms of the true minimizer β^0
$N := (1, 2, ..., p) \setminus S$ is the set of zero terms of the true minimizer β^0

So for instance X_N is the matrix of ignored predictor columns.
Card ID: 1778792585928

Thm: **Let $\lambda > 0$ and define $\Delta = X_N^\top X_S (X_S^\top X_S)^{-1}\mathbf{sgn}(\beta_S^0)$. If $\|\Delta\|_\infty \leq 1$ and for $k \in S$,**

$$|\beta_k^0| > \lambda|\mathbf{sgn}(\beta_S^0)^\top[\{\tfrac{1}{n}X_S^\top X_S\}^{-1}]_k|,$$

then there exists a Lasso solution $\hat{\beta}_\lambda^{\mathbf{L}}$ with $\mathbf{sgn}(\hat{\beta}_\lambda^{\mathbf{L}}) = \mathbf{sgn}(\beta^0)$. As a partial converse, if there exists a Lasso solution $\hat{\beta}_\lambda^{\mathbf{L}}$ with $\mathbf{sgn}(\hat{\beta}_\lambda^{\mathbf{L}}) = \mathbf{sgn}(\beta^0)$, then $\|\Delta\|_\infty \leq 1$.

Proof sketch:
Plug into the KKT conditions for the Lasso. Use the following decomposition:

$$X^{\mathrm{T}}X = \begin{bmatrix} X_S^{\mathrm{T}}X_S & X_S^{\mathrm{T}}X_N \\ X_N^{\mathrm{T}}X_S & X_N^{\mathrm{T}}X_N \end{bmatrix}$$

Also note that if $\mathrm{sgn}(\beta^0) = \mathrm{sgn}(\hat{\beta})$, then $\hat{\beta}_N = 0$.
Card ID: 1779140587978

TODO compatability condition stuff

3.6 Concentration Inequalities II

Def: Bernstein's Condition

A random variable X fulfills the Bernstein condition with parameters $\sigma, b \geq 0$ if:

$$\mathbb{E}(|X - \mathbb{E}X|^k) \leq \frac{1}{2} k! \sigma^2 b^{k-2}$$

for $k = 2, 3, ...$

Card ID: 1779140587979

Characteristic Function Bound of Bernstein's Inequality

Let $W_1, W_2, ...$ be IID random variables, all with mean μ, and each satisfying Bernstein's condition with parameter (σ, b). Then we have

$$\mathbb{E}(e^{\alpha(W_i - \mu)}) \leq \exp\left(\frac{\alpha^2 \sigma^2}{2(1 - b|\alpha|)}\right)$$

for $|\alpha| < 1/b$.

Card ID: 1779140587981

Probability bound of Bernstein's Inequality

Let $W_1, W_2, ...$ be IID random variables, all with mean μ, and each satisfying Bernstein's condition with parameter (σ, b). Then we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + bt)}\right)$$

for $t > 0$.

Card ID: 1779140587983

todo proof of bernstein's inequality

Def: Product of Subgaussian RVs

Suppose W, Z be mean zero and subgaussian with parameters σ_W and σ_Z respectively. Then WZ fulfills Bernstein's condition with parameter $(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$

Card ID: 1779140587985

todo using that as a bound for compatability condition stuff

3.7 The Lasso in Practice

Def: Coordinate Descent

Best used to optimize functions of the form

$$f(x) = g(x) + \sum_{i=1}^{n} h_j(x_j)$$

Start with some initial guess $x^{(0)}$. Optimize on one component at a time, updating as you go along each cycle.

Card ID: 1778792585929

When will coordinate descent converge to a minimizer?

When $A_0 = \{x : f(x) \leq f(x^{(0)})\}$ is a compact set (with $x^{(0)}$ the starting guess)

Card ID: 1778792585931

TODO

3.8 Extensions of the Lasso

Def: The Square Root Lasso

Minimizes

$$\frac{1}{\sqrt{n}}||Y - \bar{Y}1 - X\beta||_2 + \gamma||\beta||_1$$

Card ID: 1780945173082

Def: Group Lasso Penalty

$$\lambda \sum_{j=1}^{q} m_j ||\beta_{G_j}||_2$$

where we have group partitions $G_1 ... G_j$ and we usually define m_k

Card ID: 1780945173084

4 Chapter 3: Graphical Modeling and Causal Inference

4.1 Conditional Independence Notation

Def: Conditional Independence

Two equivalent definitions for $X \perp\!\!\!\perp Y \mid Z$, for random vectors X, Y, Z.

---

1. $\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z)\mathbb{P}(Y \in B \mid Z)$

2. $\mathbb{P}(X \in A \mid Y, Z) = \mathbb{P}(X \in A \mid Z)$
(knowing Y doesn't give us any new information about X)

note that here $X \in A$ is best read as 'X lands in A'

Card ID: 1768527597945

Def: Weak Union rule for Conditional Independence

$$X \perp\!\!\!\perp Y, Z \implies \begin{cases} X \perp\!\!\!\perp Y \mid Z \\ X \perp\!\!\!\perp Z \mid Y \end{cases}$$

We can move elements from the right-hand side to the conditioning side. Also works with some other conditioning W.

Card ID: 1768527597947

Def: Contraction rule for Conditional Independence

$$\left.\begin{array}{l} X \perp\!\!\!\perp Z \\ X \perp\!\!\!\perp Y \mid Z \end{array}\right\} \implies X \perp\!\!\!\perp Y, Z$$

We can apply known independencies to the conditioning side. Also works with some other conditioning W.

Card ID: 1768527597948

Def: Intersection rule for Conditional Independence

$$\begin{cases} X \perp\!\!\!\perp Y \mid Z \\ X \perp\!\!\!\perp Z \mid Y \end{cases} \implies X \perp\!\!\!\perp Y, Z$$

Converse of the weak union rule. ONLY ALLOWED when X, Y, Z, W have a joint density, and $f_{Y,Z,W}$ is everywhere positive (need to ensure $Y \neq Z$ basically)

Card ID: 1768527597950

When are we allowed to use the intersection rule for conditional independence?

When X, Y, Z, W have a joint density, and $f_{Y,Z,W}$ is everywhere positive (need to ensure $Y \neq Z$ basically)

Card ID: 1768527597952

4.2 Graphs

Def: Topological Ordering

A permutation π of some set of nodes $[n], n \in \mathbb{N}$, is a topological ordering if:

$$\forall a, b \in [n],\ b \in \text{de}(a) \implies \pi(a) < \pi(b)$$

(Higher nodes in the DAG are sorted to the front)

Card ID: 1768433201607

Prop

Every DAG has a topological ordering.

First, prove by contradiction that there must be a node with no parents, called the source node (or else we must have a cycle, breaking our DAG assumption).
Next, build the graph up inductively from a graph with one node.
Base case: Clearly any graph with one node is a topological order.
Inductive step: We assume that ANY graph with $n-1$ nodes, G_{n-1}, has a topological order. Thus, given a graph G_n nodes, apply the lemma we proved above to get the parent node, and remove it to get a new graph G_{n-1}. By the inductive step we have a topological order π_{n-1}. Construct π_n by shifting every node in π_{n-1} down one, and putting the parent node at the top at 1. We're done.

Card ID: 1768435499090

Def: Undirected Graph Separation

Given an undirected graph $G = ([p], E)$ and disjoint sets $A, B, S \subset [p]$, S separates A and B if...

Every path from A to B contains a node in S.

Card ID: 1769639660409

4.3 Undirected Graph Models

Def: Conditional Independence Graph (CIG)

Given a distribution P on $\mathbb{R}^p$, an undirected graph $G = ([p], E)$ where, if $Z \sim P$,

$$(j,k),(k,j) \in E \iff Z_j \not\perp\!\!\!\perp Z_k \mid Z_{-jk}$$

where Z_{-jk} is a subvector of Z without indices j and k.

Thus components are connected if there exists some dependence between them (even given all other information)

Card ID: 1768778759740

Def: Global Markov Undirected Graphs

A distribution P on $\mathbb{R}^p$ is global Markov with respect to an undirected graph $G = ([p], E)$ when...

Taking any $A, B, S \subset [p]$ with S separating A and B, we have that for any $Z \sim P$,

$$Z_A \perp\!\!\!\perp Z_B \mid Z_S$$

Card ID: 1769639660412

Thm: If a distribution P on $\mathbb{R}^n$ has a density with respect to the product measure which is everywhere positive, then its conditional independence graph is global Markov

Proof sketch: Use backwards induction on the size of the separating set. Base case: A has one element, B has one element, everything else is in S (case $p-2$). Fulfills global Markov by independence graph definition. Inductive step: Grow A and B to include all 'independent' points. Either A or B must have at least two points; WLOG it's A. Call one of the points j, and the rest of A A_-. By the inductive hypothesis we have:

$$Z_{A_-} \perp\!\!\!\perp Z_B \mid Z_{S \cup \{j\}}$$

$$Z_j \perp\!\!\!\perp Z_B \mid Z_{S \cup A_-}$$

So, applying the intersection rule (and using our assumption), we have

$$Z_A \perp\!\!\!\perp Z_B \mid Z_S$$

as required. Card ID: 1779140587992

Def: Markov Blanket

In a regression problem where X is a set of predictors and Y is a response, a Markov blanket X_S is a subset of X that fulfills

$$Y \perp\!\!\!\perp X_{S^c} \mid X_S$$

(so X_S contains all predictors relevant to predicting Y).

There's guaranteed to be a unique minimal blanket (take the weak union of every individually dependent covariate)

Card ID: 1779140587993

4.4 Directed Graphs and Causality

Def: Structural Causal Models (SCMs)

A tuple $SS = (P_j, h_j, Q_j)_{j \in [p]}$ where:
P_j is the set of 'dependencies' for node j: in other words, $\text{Pa}(j)$.
$h_j : \mathbb{R}^{|P_j|} \times \mathbb{R} \to \mathbb{R}$: A 'decision function' for node j, which takes in every defined dependency and a source of randomness, then makes a decision as to node j's value.
Q_j: A probability distribution over $\mathbb{R}$, the source of randomness for each node.
Provided alongside it is the full random vector $Z_j := h_j(Z_{P_j}, \varepsilon_j)$

Card ID: 1779140587995

Def: Interventions and Perfect Interventions to a Structural Causal Model (SCM)

Replace a $h_j(Z_{P_j}, \varepsilon_j)$ with a new $\tilde{h}_j(Z_{P_j}, \varepsilon_j)$, hoping to achieve some downstream result. If $\tilde{h}_j$ is just some constant function, we say it is a perfect intervention.

Card ID: 1779140587996

How do we denote an intervention in an SCM?

$$...|\text{do}(Z_j = \cdot)$$

Card ID: 1779140587998

Def: Collider in a path in a DAG

Given some DAG with an undirected path $j_1, ..., j_m$ with length $m \geq 3$, we say j_l is a collider (relative to the path) if we have

$$j_{l-1} \to j_l \leftarrow j_{l+1}$$

Card ID: 1779140588000

Def: Blocked path in a DAG

A path $j_1, ..., j_m$ is blocked by a set of nodes S if there exists a node v on the path such that either:
v IS NOT a collider, and $v \in S$, or
v IS a collider, and neither v nor any of its descendants are in S.

So in other words, a path is not blocked if non-colliders are not in S and colliders are either in S, or have descendants in S.

Card ID: 1779140588001

Def: d-Seperation in a DAG

and how it's denoted

Given disjoint subsets A, B in a graph G, we say A and B are d-separated by a subset S if every path between A and B is blocked by S. We denote this as:

$$A \perp\!\!\!\perp_G B \,|\, S$$

Card ID: 1779140588003

Def: Global Markov Directed (Causal) Graphs

A distribution P on $\mathbb{R}^p$ is global Markov with respect to a DAG G on p vertices if, whenever $Z \sim P$, and A, B, S are disjoint sets of vertices,

$$A \perp\!\!\!\perp_G B \,|\, S \implies Z_A \perp\!\!\!\perp Z_B \,|\, Z_S$$

Note that distributions generated from an SCM are always global Markov with respect the SCM DAG.

Card ID: 1779140588005

4.5 Practical Graphical Modeling

Prop

Let $p \geq 2$ and $Z \sim P$. If $Z_j \not\perp\!\!\!\perp Z_k \mid Z_{-jk}$ for $j, k \in [p]$, then any causal DAG generating P must have either $j \to k$, $k \to j$, or $j \to l \leftarrow k$ for some l.

Proof sketch:
By definition, j and k can not be d-seperated by $S := [p] \setminus \{j, k\}$. The only possible way to avoid being blocked by every other point is to either have a path directly between j and k, or to have $j \to l \leftarrow k$.

Card ID: 1779140588007

Prop

If nodes j and k in a DAG with p vertices are not adjacent and π is a topological order on $[p]$ with $\pi(j) < \pi(k)$, then they are d-separated by Pa(k).

Proof sketch:
Consider a path $j = j_1, ..., j_m = k$. We can't have $k \to j$ by the topological order assumption. We can't have $j_{m-1} \to k$ because then $j_{m-1} \in \text{Pa}(k)$. We could try to have a V-shaped path

$$j = j_1 \to ... \to j_{l-1} \to j_l \leftarrow ... \leftarrow j_m = k$$

(must exist due to topological ordering). But if we take the maximal j_l, we can see that j_l can't have a descendant in $\text{Pa}(k)$ because we'd have a cycle in a DAG. j and k must be d-separated.

Card ID: 1779140588009

Def: Fundamental result of Gaussian causal models

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

For nodewise regression, specialize this to $A = \{k\}$ and $B = A^c$.

Card ID: 1779140588010

Def: Graphical Lasso

Minimizes

$$-\log\det(\Omega) + \operatorname{tr}(S\Omega) + \lambda \sum_{j,k} |\Omega_{jk}|$$

where Ω is the precision matrix Σ^{-1}, and $S := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} - \hat{X}\hat{X}^{\mathrm{T}}$

Card ID: 1779140588012

4.6 Conditional Independence Testing

Def: Weak Law of Large Numbers

If $W_1, W_2, ...$ are iid real-valued random variables and $\mathbb{E}(W_i) = \mu < \infty$, then as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} \mu$$

Card ID: 1779140588013

Def: Objective of Conditional Independence Testing

Given n iid copies (x, y, z), ..., $(x_n, y_n, z_n) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$ collected into $(X, Y, Z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n \times p}$, we wish to test the null hypothesis $X \perp\!\!\!\perp Y \mid Z$.

Card ID: 1769639660415

Note that we have

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\implies \mathbb{E}\left[(X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z)) \mid Z\right] = 0 \\ &\implies \mathbb{E}\left[(X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z))\right] = 0 \end{aligned}$$

Intuitively, how can we think about $\mathbb{E}(X \mid Z)$, and how could we solve for such a thing?

$\mathbb{E}(X \mid Z)$ represents our 'best guess' for X given some information Z.

We can solve for $\mathbb{E}(X \mid Z)$ by finding an f that minimizes

$$\mathbb{E}\left[(X - f(Z))^2\right]$$

(this is called the L2 projection theorem, and it links together statistical regression and the conditional expectation)

Card ID: 1769639660417

Def: Generalized Covariance Measure (GCM)

Assuming we have data (X_i, Y_i, Z_i) and fitted regression functions $\hat{f}(z)$ and $\hat{g}(z)$...

The test statisitc

$$T := \sqrt{n}\frac{\tau_N}{\tau_D}$$

where

$$\tau_N := \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i \hat{\xi}_i$$

$$\tau_D^2 := \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 \hat{\xi}_i^2 - \tau_N^2$$

where we've defined $\hat{\epsilon}_i := X_i - \hat{f}(Z_i)$ and $\hat{\xi} := Y_i - \hat{g}(Z_i)$

Explanation:
Our null hypothesis is that

$$\mathbb{E}\left[(X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z))\right] = 0,$$

so we just need to quantify deviance from that. $\hat{f}$ and $\hat{g}$ approximately represent $\mathbb{E}(X \mid Z)$ and $\mathbb{E}(Y \mid Z)$ respectively, and so we define our average deviance among samples as τ_N.
The CLT tells us $\sqrt{n}\tau_N$ should be Gaussian (mean-zero under the null, iid per i), but we need to normalize for variance if we want a standard normal distribution. So we define τ_N to do this, using $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Card ID: 1769639660418

Thm: The Generalized Covariance Measure is asymptotically normal

Assuming there exists $c > 0$ such that $\mathrm{Var}(\epsilon \mid Z), \mathrm{Var}(\xi \mid Z) < c$ almost surely, $\mathrm{Var}(\epsilon\xi) > 0$,
and

$$\mathcal{E}_f \to 0,\ \mathcal{E}_g \to 0,\ \mathcal{E}_f\mathcal{E}_g = o(n^{-1}),$$

given $\mathcal{E}_f := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(Z_i) - \hat{f}(Z_i))^2\right]$ and likewise for $\mathcal{E}_f$, then

$$T \sim N(0, 1)$$

Card ID: 1769639660421

4.7 Average Treatment Effect Estimation

TODO

5 Chapter 4: Multiple Testing

5.1 Introduction

Def: Multiple Testing

When we have a sequence of null hypotheses $H_1 ... H_m$ where we have found associated p-values $p_1 ... p_m$.

We let I_0 represent the indices of the true nulls where no relationship exists, with $m_0 = |I_0|$.

It's not as simple as just rejecting the small-enough p-values due to the jellybeans effect!

Card ID: 1768433201608

What does I_0 represent in multiple testing?

A set of hypothesis indices representing true nulls that should not be rejected.

Card ID: 1768433201610

Def: Family-Wise Error Rate (FWER)

$$\mathbb{P}(N \geq 1)$$

where N is the number of false rejections.

For instance the FWER of the naiive approach to multiple testing is

$$1 - (1 - \alpha)^m$$

Card ID: 1768433201612

Def: Bonferroni Correction

Reject H_i if $p_i \leq \frac{\alpha}{m}$ for desired α level.

Card ID: 1768433201614

Prop

Bonferroni Correction yields a FWER $\leq \alpha$

Just apply Markov's inequality and chug:

$$\begin{aligned}
\mathbb{P}(N \geq 1) &\leq \mathbb{E}(N) \\
&\quad \text{by the Markov inequality} \\
&\leq \mathbb{E}\left[\sum_{i \in I_0} \mathbb{1}_{\{p_i \leq \alpha/m\}}\right] \\
&\leq \sum_{i \in I_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{m}\right) \\
&\leq \frac{\alpha m_0}{m} \\
&\leq \alpha
\end{aligned}$$

Card ID: 1768435499095

5.2 Closed Testing and Holm's Procedure

Def: Intersection Hypothesis

When we want to test some subset of our multiple hypotheses $I \subseteq [m]$ are all true: H_I.

We can reject it by showing that any of the hypotheses can be rejected!

Card ID: 1768433201616

Def: Local Test

and its critical requirement

Given some intersection hypothesis H_I, a test ϕ_I that is 1 if we reject H_I or zero otherwise. We require that the error rate of ϕ_I be $\leq \alpha$.

Card ID: 1768433201617

Def: Closed Testing Procedure

Suppose you have a local test ϕ_I for testing intersection hypotheses. Reject H_I iff all superset local tests ϕ_J, $I \subseteq J$ are rejected (test everything above I).

Card ID: 1768433201619

Prop

When all hypotheses are tested using closed testing, a false rejection will be made with at most probability α

False rejections can only be made if our intersection hypothesis is entirely composed of indices in I_0. But then including upwards, we are guaranteed to check ϕ_{I_0}, which only happens with probability at most α.

It follows that the FWER for any particular H_I has at most level α.

Card ID: 1768435499099

So, we can pick different local tests to use with closed testing!

Def: Holm's Procedure

Closed testing with the Bonferroni test as the local procedure:

$$\phi_I(x) = \begin{cases} 1 & \min_{i \in I}(p_i) \leq \frac{\alpha}{|I|} \\ 0 & \text{otherwise} \end{cases}$$

Card ID: 1768433201621

Sorting our p-values smallest to largest, we can message this definition into something a bit more interpretable:

Def: Holm's Procedure as a Stepdown Procedure

Let $p_{(i)}$ represent the ith smallest p-value for respective hypothesis $H_{(i)}$.

Step 1: if $p_{(1)} \leq \frac{\alpha}{m}$, reject $H_{(1)}$ and continue; otherwise accept $H_{(1)}...H_{(m)}$
...
Step i: if $p_{(i)} \leq \frac{\alpha}{m-i+1}$, reject $H_{(i)}$ and continue; otherwise accept $H_{(i)}...H_{(m)}$
...
Step m: if $p_{(m)} \leq \alpha$, reject $H_{(m)}$ and finish; otherwise accept only $H_{(m)}$

So, to recap, iterate over p-values starting at the smallest, rejecting and continuing if $p_{(i)} \leq \frac{\alpha}{m-i+1}$ or ceasing otherwise.

Card ID: 1768433201623

How good is Holm's procedure?

Uniformly more powerful than Bonferroni correction

Card ID: 1768433201624

5.3 False Discovery Rate

In some large situations, we care less about having everything correct (or rather not wrong) and more about how much we are able to reject. Need a way of measuring 'adaptiveness' to situations with a lot of signal. So, many modern approaches aim to control False Discovery Rate (FDR):

Def: False Discovery Rate (FDR)

$$\mathbb{E}\left(\frac{N}{R \vee 1}\right)$$

where N is the number of false rejections, R is the total number of rejections, and $\vee$ is the max function (so we never divide by zero).

Card ID: 1768433201626

Def: Benjamin-Hochberg Procedure

Order hypotheses by their p-values, from smallest to largest. Find the largest index underneath the slope $p_{(i)} \leq \frac{\alpha i}{m}$: that index and everything below it is rejected, but everything above i is accepted. The FDR is guaranteed to be kept at $\frac{\alpha m_0}{m} \leq \alpha$

Visual idea: find the largest index point under the line with slope $\frac{\alpha}{m}$, that is your cutoff point.

Card ID: 1768433201627

What is the benefit of using the Benjamin-Hochberg Procedure?

It's guaranteed to keep the FDR beneath α:

$$\frac{\alpha m_0}{m} \leq \alpha$$

Card ID: 1775594408152

Prop

Benjamin-Hochberg keeps FDR $\leq \alpha$

We must assume that all $p_i, i \in I_0$ are independent from all the other p-values.

Begin by breaking down R then N into their possible cases, and express as a disjoint probability in two terms.

Simplifying the event in the probability term: break down the condition $R = r$ into what it means for r to be the max value. Define a new set p_{-i} where we've removed the ith p-value (shifting the indices to the left), so that they are independent of p_i. Remark these two terms are equivalent themselves to a reduced Benjamin-Hochberg trial with a modified cutoff line $\frac{\alpha}{m}$.

Plugging back in the rewritten event, seperate the probability block and simplify to get $\leq \frac{\alpha m_0}{m} \leq \alpha$.

Card ID: 1768435499105

6 Misc

Def: The Gamma Integral

$$\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$$

Card ID: 1778890950687

Def: How to compute $\Gamma(n)$

Know that

$$\Gamma(n) = (n-1)!$$
$$\Gamma(1/2) = \sqrt{\pi}$$
$$\Gamma(z+1) = z\Gamma(z)$$

for $n \in \mathbb{N}$, $z \in \mathbb{R}$.

Card ID: 1778890950688

Def: Laplace Transform of a Power

$$t^{-\alpha} = \frac{1}{\Gamma(z)} \int_0^\infty v^{z-1} e^{-tz}$$

Card ID: 1778890950689