JSTARLING@UTEXAS.EDU
512.510.0476

Basic Terminology overview

- Model type:
  - If X contains binary indicators, ANOVA
  - If X contains continuous covariates, Regression
  - If X is a mix, 'linear model'.

- Column space of X: C(X), the span (set of all linear combos) of the cols of X.
- basis: a linearly indep. spanning set.

  ⊛ If M is a ppm onto C(X), and $v \in C(X)$, can write $Mv = v$.
  If $v \in C(X)$, can write $v = Xb$; v is a linear combo of cols of X.

- vector space: $M \in \mathbb{R}^n$ is a vec space if closed under addition & mult.
  - $\forall x, y \in M$, $(x+y) \in M$
  - $\forall x, y \in M$, $x'y \in M$
  - $N \subseteq M$ is a subspace of M if N is also a vector space.

- linearly indep: $\{x_1 \dots x_n\}$ are linearly dependent if at least one can be written as a linear combo of others.

- orthogonality: • $X \perp Y$ iff $x'y = 0$
  - $\{x_1 \dots x_p\}$ is an orthogonal basis if $x_i \perp x_j$, ie all elements orthog

- orthonormal basis: An orthog basis, but all elements have length 1.

- Gram-Schmidt: Algorithm to get an orthonormal basis out of $\{x_1 \dots x_n\}$ vecs.

  ⊛ Key: Can decompose ANY vector x as $x = x_0 + x_1$, where $x_0 \in C(X)$, $x_1 \in C(X)^\perp$
  This decomposition is unique.
  Think of as an element in x plus a residual.

Perpendicular Projection Matrices

1) Formal definition:

M is a ppm onto K-dim subspace C(X) if:

1. If $x \in C(X)$, then $Mx = x$ (leaves elems of C(X) unchanged)
2. If $x \in C(X)^{\perp}$ then $Mx = 0$ (zeros out elems orthog to C(X))

2) Constructive Definitions:

1. Let O be an orthonormal basis of C(X). (Can always obtain O via Gram-Schmidt.) Then $M = OO^T$.

2. $M = X(X'X)^{-1}X'$ is a ppm to C(X).
   $(I-M)$ is a ppm to $C(X)^{\perp}$.

3) Att. criteria for being a ppm: M is a ppm to C(X) iff

1. M is idempotent. $MM = M$
2. M is symmetric. $M = M'$

4) Rank of M: $tr(M) = r(M) = r(X)$.

Linear Algebra Review:

- A is non-singular if $\exists\ A^{-1}$ s.t. $AA^{-1} = A^{-1}A = I$.
- Rank: dimension of vec space spanned by cols of A.
  - # of linearly indep cols
  - $tr(A)$ if A idempotent.
  - A is full rank (rank = # cols) if non-singular. All ppms are full rank.

- Null space: Set of all vectors x s.t. $Ax = 0$; $N(A)$
  - If A singular, $N(A)$ = just the zero vec
  - Rank of null space = # of redundant cols: If $r(A) = r$, $r(N(A)) = n - r$ for $A_{n \times n}$

- Eigenvals/vecs: $Ax = \lambda x \longrightarrow$ x = eigvec, λ = matching eigval, for matrix A.
  - $(A - \lambda I)x = 0$ is another arrangement
  - multiplicity of an eigval = # of times it occurs
  - Distinct eigvals' eigvecs are orthogonal.

- Positive Semi-Definite: Cov matrices must be psd. (pd is ok too)
  - A is pos def (pd) if all eigvals λ are > 0.
  - A is pos semi def (psd) if all eigvals λ are ≥ 0.

- Or, A is psd if $\exists$ Q s.t. $A = QQ^T$. Useful for decomposing cov matrices.

- Singular Value Decomp (SVD):
  - $X_{(n \times p)} = UDV^T$ where
    - $D = diag(\lambda_j)$, where λ_j^2 are eigvals of $X'X$
    - $V_{p \times p} = [v_1 \ldots v_p]$ with v_j = eigvecs of $X'X$, ie $X'XV = D^2V$
    - $U_{n \times p} = [u_1 \ldots u_p]$ with u_j = eigvecs of XX', ie $XX'U = D^2U$
  - u_j's and v_j's are orthonormal

- If X symmetric, reduces to $X = PDP^T$ where $D = diag(\lambda_j)$, λ_j = eigvals of X and P = orthonorm eigvecs of X.
- Also called 'eigval decomp' or 'spectral decomp'.

- Trace Properties: $tr(A)$ = sum of diag elements
  - For A symmetric, $tr(A)$ = sum of eigvals
  - Cyclic movement of inside trace: $tr(ABC) = tr(CAB) = tr(BCA)$
  - $tr(M) = r(M)$ for ppm M.

* For an idempotent matrix, only possible eigvals are 0 and 1.

LM

Chapter 1 Theorems:

1.1.1: For A, B constant & X, Y random ~~matrices~~ vectors,
$\text{Cov}(AX+BY) = A\,\text{Cov}(X)A' + B\,\text{Cov}(Y)B' + A\,\text{Cov}(X,Y)B' + B\,\text{Cov}(Y,X)A'$

Quadratic Forms:

1.3.2: If $E(Y)=\mu$ and $\text{Cov}(Y)=\Sigma$, then $E(Y'AY) = \text{tr}(A\Sigma) + \mu'A\mu$
(Expectation of Quadratic Forms)

1.3.3: If $Y \sim N(\mu, I)$ and M is a ppm, then $Y'MY \sim \chi^2(r(M), \mu'M\mu/2)$

1.3.6: If $Y \sim N(\mu, V)$ then $Y'AY \sim \chi^2(\text{Tr}(AV), \mu'A\mu)$ under these conditions:
1. $VAVAV = VAV$
2. $\mu'AVA\mu = \mu'A\mu$ } where $AVA = A$ in all these
3. $VAVA\mu = VA\mu$

1.3.7: If $Y \sim N(\mu, \sigma^2 I)$ and $BA=0$, then
1. $Y'AY$ & BY are indep (A symmetric)
2. $Y'AY$ and $Y'BY$ are indep (A, B symmetric)

2.1 - Identifiability + Estimability

- Consider the linear model: $Y = X\beta + e$, $E(e)=0$.

- A parameterization for $E(Y)$ means writing $E(Y)$ as a fctn of some params β.

  - $E(Y) = f(\beta)$
  - A general linear model is a parameterization; $E(Y) = X\beta$

  - Why? $E(Y) = E(X\beta + e) = X\beta + E(e) = X\beta + 0 = X\beta$

- A parameterization is identifiable if knowing $E(Y)$ tells you β.

- Identifiability:
  - β is identifiable iff for any $\beta_1 \neq \beta_2$, $f(\beta_1) = f(\beta_2) \rightarrow \beta_1 = \beta_2$.
  - Fctn $g(\beta)$ identifiable iff for any $\beta_1 \neq \beta_2$, $f(\beta_1) = f(\beta_2) \rightarrow g(\beta_1) = g(\beta_2)$.
  - Ex: If $X\beta_1 = X\beta_2$, then $\beta_1 = (X'X)^{-1}X'X\beta_1 = (X'X)^{-1}X'X\beta_2 = \beta_2$.

- Thm 2.1.2: A fctn $g(\beta_1)$ is identifiable iff $g(\beta)$ is a fctn of $f(\beta)$.

- Estimability
  - An estimable fctn is an identifiable fctn which is linear.
  - Means you can write as a fctn of design matrix X.

  - A scalar-valued linear fctn of β: $\lambda'\beta = p'X\beta$ for vector p.
  - A vector-valued linear fctn of β: $\Lambda'\beta = P'X\beta$ for matrix P.

  - p and P are not unique, but MP is (and Mp), ie projection of p / P onto $C(X)$.

- Linear Estimate: $f(Y)$ is a linear est of $\lambda'\beta$ if $f(Y) = a_0 + a'Y$ for scalar a_0, vec a,

  - $\hat{\theta}$ is a linear est of θ if has form $\hat{\theta} = a_0 + a_1'\tilde{\theta}$

- Biased Estimators:
  - $\hat{\theta}$ unbiased if $E(\hat{\theta}) = \theta$
  - $a_0 + a_1'Y$ is unbiased est of $\lambda'\beta$ iff $a_0 = 0$ and $a_1'X = \lambda'$

2.2 Least Squares Estimation

- Model: $Y = X\beta + e$, $E(e)=0$, $Cov(e)=\sigma^2 I$
- $\hat{\beta}$ is a least sqs (LS) est if it minimizes $\min_{\beta}\{(Y-X\beta)'(Y-X\beta)\}$
- $\hat{\beta} = (X'X)^{-}X'y$, and $\hat{y} = X\hat{\beta} = My$ where M is ppm to $C(X)$: $M = X(X'X)^{-}X'$
- For estimable fctn $\lambda'\beta = p'X\beta$, $\lambda'\hat{\beta} = p'My$
  - If $\lambda' = p'X$, then $E(p'My) = \lambda'\beta$, ie $p'My$ unbiased.
- Thm 2.2.6: Let $r = r(X)$ and $Cov(e)=\sigma^2 I$. Then $\hat{\sigma}^2 = \frac{Y'(I-M)Y}{n-r}$ is unbiased est. of σ^2.
  - $\hat{\sigma}^2 = \frac{Y'(I-M)Y}{n-r} = \frac{RSS}{n-r}$ RSS = squared length of residual vector.
  - RSS is also called SSE; sum of squared error
  - $\hat{\sigma}^2$ is also called MSE; mean squared error.
  - $(n-r) = \text{rank}(I-M) = \text{df error}$

2.3 Best Linear Unbiased Estimation (BLUE)

- "Best" = smallest variance.
- $a'Y$ is BLUE for $\lambda'\beta$ if
  - $E(a'Y) = \lambda'\beta$
  - $\text{Var}(a'Y) \le \text{Var}(b'Y)$ for any other unbiased $b'Y$.
- Gauss-Markov Theorem: If $\lambda'\beta$ estimable, then $\lambda'\hat{\beta}_{LS}$ is BLUE.
  (If $\lambda'\beta = p'X\beta = X\beta$, $\hat{\beta}_{LS}$ is BLUE for β.) → Above is more general version.
- Proof: Let M be the ppm to $C(X)$. $\lambda'\beta = p'X\beta$, so LS est of $\lambda'\beta$ is $\lambda'\hat{\beta}_{LS} = p'My$.
  Let $a'y$ be another unbiased estimator of $\lambda'\beta$.
  Show $\text{var}(a'y) \ge \text{var}(p'My)$.

$$\text{Var}(a'Y) = \text{Var}(a'Y - p'MY + p'MY)$$
$$= \underbrace{\text{Var}(a'Y)}_{\ge 0} + \text{Var}(p'My) - \underbrace{2\,\text{Cov}(a'y - p'My,\ p'My)}_{=0}$$

Why does $\text{Cov}(a'y - p'My,\ p'My) = 0$?

$$\text{Cov}(a'y - p'My,\ p'My) = (a' - p'M)\text{Cov}(y)\,Mp$$

Then since $a'y$ unbiased, $\lambda'\beta = E(a'y) = a'X\beta$, so $p'X\beta = a'X\beta$, ie $p'X = a'X$.

Then $M = X(X'X)^{-}X'$, so have:

$$(a' - p'M)\,\text{Cov}(y)\,Mp = \sigma^2(a' - p'M)Mp = \sigma^2(a'M - p'M)p$$, and from above,

$a'M = p'M$, so $= 0$. ∎

Corollary 2.3.3: If $\sigma^2 > 0$, ∃ a unique BLUE for any estbl fctn $\lambda'\beta$.

2.4 Max Lhood Estimation

- Assume $Y \sim N(X\beta, \sigma^2 I)$. Then can estimate $\hat{\beta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ by maximizing

$$L = (\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right]$$

- For convenience, maximize $\log L$. Gives same result as LS for $\hat{\beta}$: $\hat{\beta} = (X'X)^{-}X'y$
- $\hat{\sigma}^2_{MLE} = \frac{Y'(I-M)Y}{n}$ is rarely used: MSE is used for $\hat{\sigma}^2$ bc unbiased.

2.6 Sampling Distributions of Estimators

- $Y \sim N(X\beta, \sigma^2 I)$
- $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-})$
- For $\lambda'\beta$, least sqs est is $\lambda'\hat{\beta} = p'MY$. Can write sampling dist in 2 ways:

$$\lambda'\hat{\beta} \sim N(\lambda'\beta,\ \sigma^2\lambda'(X'X)^{-}\lambda)$$

$$p'My \sim N(\lambda'\beta,\ \sigma^2 p'Mp)$$

2.7 Generalized Least Squares:

- Model: $Y = X\beta + e$, $E(e)=0$, $Cov(e) = \sigma^2 V$ where V = psd, known. (1)
- Write $V = QQ^{-1} \longrightarrow Q^{-}VQ^{-T} = I$, so left-multiply and solve new model:

$$\underbrace{Q^{-1}Y}_{Y^*} = \underbrace{Q^{-1}X}_{X^*}\beta + \underbrace{Q^{-1}e}_{e^*} \qquad \text{with } E(e^*)=0,\ Cov(e^*)=\sigma^2 I \qquad (2)$$

- Called Gen LS bc no longer minimizing squared dist btwn $Y = X\beta$, but minimizing a generalized squared distance specified by V^{-1}.

LM

Thm 2.7.1:

(a) $\lambda'\beta$ is estbl in model (1) iff estbl in model 2.

(b) $\hat{\beta}$ is the gls est iff $\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}y$

· for any estimable fctn, $\exists$ a unique LS soln.

(c) $\lambda'\hat{\beta}$ is the BLUE of $\lambda'\beta$.

(d) If $e \sim N(0, \sigma^2 V)$ then $\lambda'\hat{\beta}_{LS}$ is the min. var unbiased estimator.

(e) If $e \sim N(0, \sigma^2 V)$ then $\lambda'\hat{\beta}_{LS}$ is also the MLE.

Proof:

(a) If $\lambda'\beta$ estbl in (1), can write $\lambda' = p'X = (p'Q)Q^{-1}X$ so $\lambda'\beta$ estbl in (2).
If $\lambda'\beta$ estbl in (2), can write $\lambda' = p'Q^{-1}X = (p'Q^{-1})X$ so $\lambda'\beta$ estbl in (1).

(b) $(X'Q^{-T}Q^{-1}X)^{-}X'Q^{-T}Q^{-1}y = Q^{-1}X\hat{\beta} \longrightarrow (X'V^{-1}X)^{-}X'V^{-1}y = \hat{\beta}$

(c) $\lambda'\hat{\beta}$ is BLUE for $Q^{-1}y$ by Thm 2.2.3. Any fctn of y can be obtained through $Q^{-1}y$ bc Q^{-1} invertible. So the gls est is BLUE.

2.8 - Normal Equations:

Solving for β gives same $\hat{\beta}$ as LS.

· $X'X\beta = X'y$ · $X'V^{-1}X = X'V^{-1}y$ for gen LS

Regression in canonical form + ridge regression are two techniques to deal w. collinearity - when cols of X are nearly linearly dependent, ie very correlated.

15.2 - Regression in Canonical Form

· gives a convenient linear transformation so can read off LS soln.

· Model: $Y = X\beta + e$

· First, need SVD:

· Thm 15.1.2: SVD (Singular Value Decomp)

· Let $X_{n\times p}$ be full rank. Can write $X = UDV^T$, where

· $D_{p\times p}$ = diag(λ_j) where λ_j are sqrts of eigvals of $X'X$ (λ_j^2 = eigvals)

· $U_{n\times p}$ = eigvecs of XX', so $X'XU =$ UD^2

· $V_{p\times p}$ = eigvecs of $X'X$, so $X'XV =$ VD^2

· Canonical form:

· Begin with $Y = X\beta + e$ and write $X = UDV^T$

· write $U_* = [U, U_1]$ where cols of U_* are an orthonormal basis for $\mathbb{R}^n$.

· Left-multiply by U_* to transform model. $\underbrace{U_*'Y}_{Y_*} = U_*'X\beta + \underbrace{U_*'e}_{e_*}$

· Then model is $Y_* = U_*'X\beta + e_*$, $E(e_*) = 0$, $Cov(e_*) = \sigma U^{-1}U_* = \sigma^2 I$

· Then use SVD again on $U_*'X\beta$ term:

$$U_*'X\beta = U_*'UDV^T\beta = \begin{bmatrix} U' \\ U_1' \end{bmatrix} UDV'\beta = \begin{bmatrix} I \\ 0 \end{bmatrix} V'\beta$$

· Reparameterize by letting $\gamma = V'\beta$, to get canonical model:

$$Y_* = \begin{bmatrix} I \\ 0 \end{bmatrix}\gamma + e_*, \quad E(e_*) = 0, \quad Cov(e_*) = \sigma^2 I$$

· Parameter estimation in this model:

· $\hat{\gamma} = D^{-1}U'y$

· $Cov(\hat{\gamma}) = \sigma^2 D^{-2}$

· $RSS = Y_*' \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix} Y_*$

15.3 - Ridge Regression:

- originally proposed to deal w. collinearity. Now viewed as form of penalized lhood estimation. (Bayesian)
- ℓ_2-norm penalty
- Shrinks coefficients, but none all the way to zero.

- loss fcn: $\min_\beta \left\{ (y-X\beta)'(y-X\beta) + \lambda\beta'\beta \right\}$ i.e. $\min_\beta \left\{ \sum_{i=1}^n (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$

- MSE intuition: $MSE = E\left[(\hat\beta-\beta)^T(\hat\beta-\beta)\right] = tr\left[\sigma^2(X'X)^{-1}\right]$ since $E(x'Qx) = tr(Q\Sigma) + \mu'Q\mu$
  - $E(\hat\beta-\beta) = \mu = 0$
  - $Q = I$
  - $\Sigma = cov(\hat\beta)$

  $= \sigma^2 \sum_{j=1}^p \frac{1}{\lambda_j^2}$ where λ_j^2 = eigvals

  - If 1+ eigvals too close to 0, MSE blows up.
  - For RR, mse becomes $\sigma^2 \sum_{j=1}^p \frac{1}{\lambda_j^2+\lambda}$, preventing denoms from being too small.

- Bayesian Interpretation:
- RR can be interpreted as a posterior mean w. a normal prior.
  - Lhood: $p(y|\beta) \sim N(X\beta, \sigma^2 I)$
  - Prior: $p(\beta) \sim N(0, \frac{\sigma^2}{\lambda} I)$
- The normal-normal update yields posterior
  - $p(\beta|y) \sim N(m, v)$ with $v = \left[\frac{1}{\sigma^2}X'X + \frac{\lambda}{\sigma^2}I\right]^{-1}$

$$m = v \cdot X'y = \left[\frac{1}{\sigma^2}(X'X+\lambda I)^{-1}X'y\right] = \frac{\hat\beta_{RR}}{\sigma^2}$$

- LS solution: has closed form, $\hat\beta_{RR} = (X'X+\lambda I)^{-1}X'y$

How does RR handle collinearity? See SM1, HW2, Prob 5:

- Run RR on model with $X = [x_1 \dots x_p]$. The $\hat\beta_{RR_1} = a$ for x_1. If we add $(m-1)$ copies of x_1, how do the identical copy coeffs relate to a?

Model 1 Loss: $\min_\beta \left\{ \sum_{i=1}^n (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} = \min_\beta \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \beta_1 x_{i1} - \sum_{j=2}^p \beta_j x_{ij} \right) + \lambda \sum_{j=1}^p \beta_j^2 \right\}$

Model 2 Loss: Now $j=1$ to m are x_1 predictors, and $j+1$ to q are $x_2 \dots x_p$.

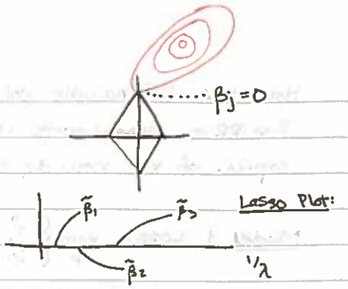
$$\min_\beta \left\{ \sum_{i=1}^n \left(y_i - \tilde\beta_0 - \sum_{j=1}^m \tilde\beta_1 x_{i1} - \sum_{j=m+1}^q \tilde\beta_j x_{ij} \right) + \lambda \left(\sum_{j=1}^m \tilde\beta_1^2 + \sum_{j=m+1}^q \beta_j^2 \right) \right.$$

- Want $\beta_1 x_{i1} = \sum_{j=1}^m \tilde\beta_1 x_{i1} \rightarrow a x_{i1} = m\tilde\beta_1 x_{i1} \rightarrow \tilde\beta_1 = a/m$.

LM

VARIABLE SELECTION METHODS

1. **LASSO:** Regression with a ℓ_1-norm penalty.
   - Shrinks some coeffs all the way to 0, unlike ridge.
   - Downside: shrinks even non-zeroed-out coeffs.

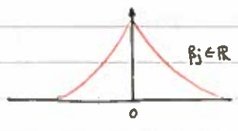


$$\hat{\beta}_L = \min_{\beta} \left\{ (y-X\beta)^T(y-X\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- No closed-form soln; solve using SGD or other optimization techniques.
- Have to tune penalty parameter λ.

2. **Bayesian Lasso:** We can do some reverse engineering to interpret $\hat{\beta}_L$ as the posterior mean of a Bayesian model

- Lhood: $p(y|\beta,\sigma^2) \sim N(X\beta, \sigma^2 I)$
- Prior: $p(\beta_j) \sim \frac{\lambda}{2\sigma} \exp\left[\frac{-\lambda}{\sigma}|\beta_j|\right]$ The 'laplace', or double exponential, prior.



- Posterior: $\log p(\beta|y,\sigma^2) = c - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta) - \frac{\lambda}{\sigma}\sum_{j=1}^{p}|\beta_j|$

- Including the highlighted σ in $p(\beta_j)$ ensures posterior is unimodal

- Note: $p(\beta_j|\sigma) = \frac{1}{2}\cdot\frac{\lambda}{\sigma}\exp\left[-\frac{\lambda}{\sigma}|\beta_j|\right] = \int N(\beta_j|0, \sigma^2\gamma_j^2)\cdot \text{Exp}\left[\gamma_j^2 \,|\, \frac{\lambda^2}{2}\right] d\gamma_j^2$

so we can replace the prior on β with:

- $p(\beta_j|\gamma_j^2,\sigma^2) \sim N(0, \sigma^2\gamma_j^2)$
- $p(\gamma_j) \sim \text{Exp}(\lambda^2/2)$

- Proof: $\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-\frac{z^2}{2s}} \cdot \frac{a^2}{2} e^{-\frac{a^2}{2}s} ds$ where $z = \beta_j/\sigma$, $a = \lambda$, $s = \gamma_j^2$

3. **Horseshoe Prior:**

L: $p(y|\beta,\sigma^2) \sim N(X\beta, \sigma^2 I)$

P: $p(\theta_i|\lambda_i) \sim N(0, \lambda_i^2)$
$p(\lambda_i|\tau) \sim C^+(0,\tau)$
$p(\tau|\sigma^2) \sim C^+(0,\sigma)$
$p(\sigma^2) \sim 1/\sigma^2$ (Jeffreys)

- Shrinks some params, leaves others as is.
- Less shrinkage on strongest predictors.

- $C^+(0,\tau)$ = half-cauchy w/ scale τ

↓

$$p(x) = \frac{1}{\pi\lambda[1+\lambda^2]}$$

LM

3.1 - Testing Background

- Model: $Y = X\beta + e$, $E(e) = 0$, $Cov(e) = \sigma^2 I$
- The ppm M is the crucial item for model testing.

- $C(X)$ = estimation space, $C(X)^\perp$ = error space, so $(I-M)$ is ppm to error space.
- Any two models w. same estimation space are essentially same model.

3.2 - Testing Models

- Full model, which we assume correct: $Y = X\beta + e$, $E(e) = 0$, $Cov(e) = \sigma^2 I$. (M1)
- Want to test if a reduced model is better: $Y = X_0\gamma + \eta$, $E(\eta) = 0$, $Cov(\eta) = \sigma^2 I$. (M0)

- $C(X_0) \subseteq C(X)$

- Test statistic:
$$\frac{Y'(M_1 - M_0)Y \;/\; \overbrace{r(M-M_0)}^{df_1}}{Y'(I-M_0)Y \;/\; \underbrace{r(I-M_1)}_{df_0}} \sim F(df_1, df_0, ncp)$$

- ncp: $ncp = \beta' X'(M-M_0)X\beta \,/\, (2\sigma^2)$

  - $ncp = 0$ under M_0, since $(M-M_0) = M(I-M_1)$, and $I-M_0 \in C(X_0)^\perp \subseteq C(X)^\perp$, so $M(I-M_0) = 0$.

- Null Hypothesis: H_0: M_0 is sufficient, reject for large F stat.

- Proof of distribution of test stat:

  1) $Y'(M_1-M_0)Y \perp\!\!\!\perp Y'(I-M_0)Y$, since by Thm 1.3.7, if $(M_1-M_0)(I-M_1) = 0$, are indep. and $(M_1-M_0)(I-M_1) = (M_1-M_0)\underbrace{M_1(I-M_1)}_{=0} = 0$

  2) Numerator: $Y'(M_1-M_0)Y \longrightarrow$ First, $Y \sim N(X\beta, \sigma^2 I)$ so $Y/\sigma^2 \sim N(X\beta, I)$. Then by Thm 1.1.3, $Y'\frac{(M_1-M_0)}{\sigma^2}Y \sim \chi^2\left[r(M-M_0), \frac{\beta'X'(M-M_0)X\beta}{2\sigma^2}\right]$

  3) Denominator: $Y'(I-M_1)Y \longrightarrow$ By Thm 1.1.3, $Y'(I-M_1)Y \sim \chi^2\left[r(I-M), \frac{\beta'X'(I-M)X\beta}{2\sigma^2}\right]$

We can set up a reduced model by specifying estimable constraints, $\lambda'\beta = 0$.
But sometimes this doesn't work, like for $H_0: \beta_1 = \beta_2 + 5$. Need a generalized testing procedure.

3.2.1 - Generalized Test Procedure:

- $M_1 = Y = X\beta + e$
- $M_0 = Y = X_0\gamma_0 + \eta + Xb$ where b = constant

- **Procedure:** $M_1: \underbrace{Y - Xb}_{y^*} = \underbrace{(X\beta - Xb)}_{X\beta_*} + e \longrightarrow M_1: y_* = X\beta_* + e$ where $\beta_* = \beta - b$

  - $M_0: y^* = X_0\gamma^* + \eta$
  - Then F-test like usual.

- **Example:** $M_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$

  - Want to test $H_0: \beta_1 = 0$, & $\beta_2 = \beta_3 + 5$

  - M_0: $y = \beta_0 + (\beta_3 + 5)x_2 + \beta_3 x_3 + e$
    $y = \beta_0 + \beta_3(x_2 + x_3) + 5x_2 + e$
    ~~$y = \beta_0 + \beta_3(x_2 + x_3) + 5x_2 + e$~~

$$(y - 5x_2) = (\beta_0 - 5x_2) + (\beta_3 - 5x_2)(x_2 + x_3) + e$$

Numerator for F-Test: $(y - 5x_2)^T (m - m_0)(y - 5x_2)$

3.3 - Testing Linear Parametric Fctns:

- M_1: $Y = X\beta + e$
- M_0: $Y = X\beta + e$, subject to $\lambda'\beta = 0$, where $\lambda'\beta$ estbl, s.t. $\lambda'\beta = P'X\beta$.

- Pick a matrix U so $C(U) = C(\lambda)$, ie U s.t. $\lambda U = 0$.
- Set $X_0 = XU$
- Then constraint is $\beta = U\gamma \longrightarrow$ M_0 is now $Y = X_0\gamma + \eta$

* Number cols for U: $\dim(U) = \dim(X) - \dim(\lambda)$

- **Example: 1-Way ANOVA**

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}_{6\times 1} = \begin{array}{c} \begin{matrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

- So $\beta = [\mu, \alpha_1, \alpha_2, \alpha_3]^T$
- Want to test $\lambda' = (0, 1, 0, -1)$
- Can pick $U = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$

so that $\lambda'U = (0\ 1\ 0\ -1)\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = (0\ 0\ 0)$ ✓

- Can pick any $U \rightarrow$ y not so easy to see, use Gram-Schmidt.
- Now M_0: $Y = X_0\gamma + \eta$ where $X_0 = XU$

- **Example 2:** Consider joint constraint $\lambda' = \begin{bmatrix} \lambda_1' \\ \lambda_2' \end{bmatrix}$ where $\lambda_1' = (0, 1, 0, -1)$ s.t. $\lambda'\beta = 0$
  $\lambda_2' = (0, 1, 1, 1)$

- Want $\lambda'U = 0$: $\begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -2 \\ 0 & 1 \end{bmatrix}}_{U} = 0$

* cols $U = 4 - 2 = 2$

- $X_0 = XU$, and $M_0 = Y = X_0\gamma + \eta$

LM

How to get back + forth, if have full model $Y = X\beta + e$?

$M_0: Y = X\gamma + \eta$ $\quad X_0 = XU$ s.t. $\lambda'U = 0$

$M_0: Y = X\beta + e$ s.t. $\lambda'\beta = 0$

Test Stat: $y'(M - M_0)y$

Test stat $y'(M-M_0)y =$
$y'M_{MP}y$ where $M_{MP} = MP(P'MP)^- P'M$
ie the ppm to MP, where P comes from $\lambda'\beta = P'X\beta$

· Prop 3.3.2: $c(M - M_0) = c(MP)$

For $\lambda'\beta$ estimable, can write test stat in terms of λ and β'.

1. $r(\lambda) = r(M_{MP}) = r(MP)$
2. Rewrite $y'M_{MP}y$. Know $X\hat\beta = My$, so $\lambda'\hat\beta = P'X\hat\beta = P'My$.

$y'M_{MP}y = y'MP(P'MP)^- P'My = \hat\beta'\lambda(P'X(X'X)^- X'P)^-\lambda'\hat\beta$, so

$y'(M - M_0)y = \hat\beta\lambda[\lambda'(X'X)^-\lambda]^-\lambda'\hat\beta$

Test stat: $\dfrac{\hat\beta\lambda[\lambda'(X'X)^-\lambda]^-\lambda'\hat\beta \;/\; r(\lambda)}{MSE}$

3.3 - What happens if you pick a non-estimable constraint?

· Say pick M_0: $Y = X\beta + e$, $\lambda'\beta = 0$ for a non-estbl constraint.

· Call $\lambda_0'\beta$ the estimable part; then $\lambda_0'\beta$ and $\lambda\beta$ give the same reduced model, since the non-estimable part doesn't induce any constraint on $C(X)$.

· Thm 3.3.6: If $C(\lambda) \cap C(X) = C(\lambda_0)$, and $C(u_0) = C(\lambda_0)^\perp$, then $C(XU) = C(XU_0)$.
Thus $\lambda_0'\beta$ and $\lambda'\beta$ induce the same model.

· Prop 3.3.7: If $\lambda'\beta = 0$ estbl, and $\lambda \neq 0$, then $\lambda'\beta = 0 \longrightarrow C(XU) \neq C(XU_0)$.
(for an estbl constraint, there is always a restraint on $C(X)$.)

· Cor. 3.3.8: $C(\lambda) \cap C(X) = \{0\}$ iff $C(XU) = C(X)$, ie if $\lambda'\beta$ has no estbl part, the constraint doesn't affect the model.

LM

3.6- Decomposition of Sums of Squares

- Total sum of squares (SS): $SS_{Tot} = SS = Y'Y$

- SSR(X) = sum of sqs for regression on X: $SSR(X) = Y'MY$
- SSE = sum of sqs for error: $SSE = Y'(I-M)Y$

$$SS = Y'MY + Y'(I-M)Y \quad \text{(by adding/subtracting } Y'MY \text{ to } Y'Y)$$

- Say M_0 is a reduced model: $X = [X_0 \mid X_1]$ with M_0 as ppm for reduced model.
- Then we can decompose $Y'MY$:

$$\underset{SSR(X)}{Y'MY} = \underset{SSR(X_1|X_0)}{Y'(M-M_0)} + \underset{SSR(X_0)}{Y'M_0Y}$$

- Other notation: Can write in terms of parameter names, say $X = [\alpha \mid \mu_1]$
  Then $SSR(X) = R(\mu, \alpha) + R(\alpha)$

- Conceptual note: $M - M_0 = M_*$, so M_* is ppm to $C(M-M_0)$.
  - Say $R = [\frac{1}{P_1} \cdots \frac{1}{P_r}]$ is an orthonormal basis of $C(X_*)$.
  - Then $M_* = RR^T = \sum_{s=1}^{r} P_s P_s^T$
  - ↓
  - So we can write any space as an additive decomp of orthog. projection matrices. We only want to do this when M_*'s have meaning, like M_0.

When order added to model matters?

- When is $R(\alpha \mid \eta, \mu) = R(\alpha \mid \mu)$? This means α, η not confounded; order doesn't matter.

- Prop 3.6.3: $R(\alpha \mid \eta, \mu) = R(\alpha \mid \mu)$ iff $C(M_1 - M_3) \perp C(M_0 - M_3)$

- Equivalently, $\underbrace{(M_1 - M_3)}_{\text{Tests dropping } \eta \text{ from model with } \mu, \eta} = \underbrace{(M - M_0)}_{\text{Tests dropping } \eta \text{ from model with } \mu, \eta, \alpha}$ iff $(M_1 - M_3)(M_0 - M_3) = 0$

  - M_1 = model with μ, η
  - M_0 = model with μ, α
  - M = model with μ, α, η

4.1- One way ANOVA Model

- Model: $y_{ij} = \mu + \alpha_i + e_{ij}$ for $i = 1 \ldots t$, $j = 1 \ldots n_i$ (t indexes trts, j indexes obs in ea. trt)
  - μ = grand mean
  - α_i = effect of ith treatment

- Common reduced model: $y_{ij} = \mu + e_{ij}$; $H_0: \alpha_1 = \ldots = \alpha_t = 0$
- Can also test various contrasts.

- $Y = X\beta + e$, $E(e) = 0$, $Cov(e) = \sigma^2 I$

- $X = \begin{bmatrix} \mu & \alpha_1 & \alpha_2 & \alpha_3 \\ | & \mathbf{1}_{n_1} & 0 & 0 \\ | & & 0 & 0 \\ 1 & & \mathbf{1}_{n_2} & 0 \\ | & 0 & & 0 \\ | & 0 & 0 & \mathbf{1}_{n_3} \\ | & 0 & 0 & \end{bmatrix}$ and $\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$ ($n_i \times n_i$ matrix of 1's)

- $M = X(X'X)^- X'$ = block diagonal s.t. $\text{diag}\left[\frac{1}{n_i} J_{n_i}^{n_i}\right] = \begin{bmatrix} \frac{1}{n_1} & \frac{1}{n_1} & \frac{1}{n_1} & & & & & \\ \frac{1}{n_1} & \frac{1}{n_1} & \frac{1}{n_1} & 0 & & 0 & & \\ \frac{1}{n_1} & \frac{1}{n_1} & \frac{1}{n_1} & & & & & \\ & & & \frac{1}{n_2} & \frac{1}{n_2} & & & \\ 0 & & & \frac{1}{n_2} & \frac{1}{n_2} & 0 & & \\ & & & & & & \frac{1}{n_3} & \end{bmatrix}$

  (for $n_1 = 3$, $n_2 = 2$, $n_3 = 1$)
  - can't just invert $(X'X)$ since not full rank- used ginv.

- Can also derive M using $Z = [\frac{1}{\alpha}, \frac{1}{\alpha_2}, \frac{1}{\alpha_3}]$ b/c $C(X) = C(Z)$. So $M = Z(Z'Z)^{-1}Z'$ gives same result.

- Fitted values: $\hat{y} = My = [\bar{y}_1 \ \bar{y}_2 \ \ldots \ \bar{y}_t]^t$
  - logical bc model assumes same means for all obs under ea. trt.
  - $\hat{y} \in C(X)$
  - $y = \hat{y} + \hat{e}$, so residuals $\hat{e} = y - \hat{y} = y - My = (I-M)y$, so $\hat{e} \in C(X)^\perp$.

- Projection matrices:
  - $M_\mu = X(X'X)^- X'$ for $X = [\frac{1}{1}] = \frac{1}{n} J_n^n$, ie a $(n \times n)$ matrix of $1/n$.
  - $M_* = M - M_\mu$

- Then $M_* y = (M - M_3)y = (\bar{y}_1 - \bar{y}, \ldots, \bar{y}_t - \bar{y})$

- Testing for contrasts: Use same F-Test as for other $H_0: Y = X\beta + e$ s.t. $L'\beta = 0$ for estimable $L'\beta$ fctn.

(goes with bottom of next pg)

· **Decomp of SS:**

· $Y'Y = Y'MY + Y'(I-M)Y$ where $Y'MY = Y'M_\mu Y + Y'\underbrace{(M-M_\mu)}_{M_\alpha}Y$

· For ANOVA Table, need Expected Mean Square, ie $E(Y'AY)$ for each SS term.

· In general, we know $E(Y) = X\beta$, and $Var(Y) = \sigma^2 I$, and $E(Y'AY) = tr(AVar(Y)) + \mu'A\mu$

Source	Expected Mean Square: E(SS/df)	df
· $Y'M_\mu Y$	$E(Y'M_\mu Y/1) = \sigma^2 + \beta'X'M_\mu X\beta$	$r(M_\mu)=1$
· $Y'(M-M_\mu)Y$	$E(Y'M_\alpha Y/t-1) = \sigma^2 + \beta'X'M_\alpha X\beta/t-1$	$r(M-M_\mu)=t-1$
· $Y'(I-M)Y$	$E(Y'(I-M)Y/n-t) = \sigma^2 + \underbrace{\beta'X'(I-M)X\beta}_{=0 \text{ since } (I-M)X=0}/n-t = \sigma^2$	$r(I-M)=n-t$

· **ANOVA Table:**

Source	df	SS	E(MS) = E(SS/df)
Grand Mean	1	$Y'M_\mu Y$	$\sigma^2 + \beta'X'M_\mu X\beta$
Trts	t-1	$Y'(M-M_\mu)Y$	$\sigma^2 + \beta'X'(M-M_\mu)X\beta/t-1$
Error	n-t	$Y'(I-M)Y$	σ^2
Total	n	$Y'Y$	

4.2 - Estimation & Contrasts for ANOVA

· $\lambda'\beta$ where μ is not involved : $\lambda' = (0, \lambda_1, \ldots \lambda_t)$

· Need to find a contrast λ so can write as $p'X\beta$, where $p'J=0$ so 1st element of $\lambda = 0$.

· $\lambda'\hat{\beta} = \sum_{i=1}^{t} \lambda_i \bar{y}_i = p'My$

· Use same F-test as for any other estimable fctn constraint:

MO = $Y = X\beta + e$ s.t. $\lambda'\beta = 0$

· A regression model is any linear model $Y = X\beta + e$ where $X'X$ is non-singular (rank p).

6.1 - Simple Linear Regression:

· Everything we did before applies; now just deriving non-matrix forms.

· **Model:** $y_i = \beta_0 + \beta_1 x_i + e_i$, or $Y = X\beta = e$ where $X = [\frac{1}{1} \; \frac{x_i}{x_i}]$, $\beta = [\beta_0 \; \beta_1]^T$

· **LS estimates:** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ $\qquad \hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

· **Derivation with mean-centered covariates:**

· $X = [J \; x_i] \longrightarrow X_* = [J \; x_i - \bar{x}]$

· note that $J \perp (x_i - \bar{x})$, so two cols are orthog.

· Can write $X_* = XU$, where $U = \begin{bmatrix} 1 & -\bar{x} \\ 0 & 1 \end{bmatrix}$ so $[J \; x_i]\begin{bmatrix} 1 & -\bar{x} \\ 0 & 1 \end{bmatrix} = [J \; x_i - \bar{x}]$

· Model spec. is unchanged: $E(Y) = X\beta = X_*\gamma$

· Solve for γ: $X\beta = XU\gamma$, so $\gamma = U^{-1}\beta$.

· $\hat{\gamma} = (X_*^T X_*)^{-1} X_*^T y$

· $(X_*^T X_*) = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^{n}(x_i-\bar{x})^2 \end{bmatrix} \longrightarrow (X_*^T X_*)^{-1} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{bmatrix}$

· $\hat{\gamma} = \begin{bmatrix} 1/n & 0 \\ 0 & 1/\sum_{i=1}^{n}(x_i-\bar{x}) \end{bmatrix}\begin{bmatrix} -J- \\ -x_i-\bar{x}- \end{bmatrix}\begin{bmatrix} 1 \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{y} \\ SS_{xy}/SS_x \end{bmatrix}$

↑ since $\bar{y}=0$, $\Sigma(x_i-\bar{x})y_i = SS_{xy}$

LM

6.2- Multiple Regression

- Model: $Y = X\beta + e$ where $e \sim N(0, \sigma^2 I)$ $\longrightarrow$ $Y \sim N(X\beta, \sigma^2 I)$
- $\hat{\beta} = (X'X)^{-1}X'y$
- $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- $SSR(x) = Y'MY = \hat{\beta}'X'X\beta$
- $SSE = Y'(I-M)Y$
- $dfE = r(I-M) = n-p$
- $\hat{\sigma}^2 = MSE = Y'(I-M)Y / r(I-M)$
- $M = X(X'X)^{-1}X'$

- Any $\lambda'\beta$ is estimable, since X is full rank.
- $\lambda'\hat{\beta} \sim N(\lambda'\beta, \sigma^2\lambda'(X'X)^{-1}\lambda)$

- Tests + CI based on $\lambda'\beta$ are:

$$TS = \frac{\lambda'\hat{\beta} - \lambda'\beta}{\sqrt{MSE \cdot \lambda'(X'X)^{-1}\lambda}} \sim t_{dfE}$$

- SS Decomp:

$$Y'Y = Y'MY + Y'(I-M)Y$$
$$SS = SSR(x) + SSE \qquad SSR(x) = Y'(M-M_J)Y + Y'M_JY$$
$$SS_{covariates} \quad SS_{int}$$

- Alt. Mean-Centered Model:
  - $M_J = \frac{1}{n}J_n^n$, the ppm for the intercept-only model.
- $Y = [J \mid (I-M_J)Z]\begin{bmatrix}\gamma_0 \\ \gamma_*\end{bmatrix} + e$, ie $Y = X_*\gamma + e$. Note $C(X_*) = C(X)$.

- γ to β mapping: $y_i = \gamma_0 + \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\gamma_j + \eta_i$ Then distribute:

$$y_i = \underbrace{\gamma_0 - \sum_{j=1}^{p}(\bar{x}_j\gamma_j)}_{\beta_0} + \underbrace{\sum_{j=1}^{p}x_{ij}\gamma_j}_{\text{for } j=1 \text{ to } p,\ \gamma_j = \beta_j} + \eta_i$$

- Normal Eqns:

$$\begin{bmatrix} n & 0 \\ 0 & Z'(I-M_J)Z \end{bmatrix}\begin{bmatrix}\gamma_0 \\ \gamma_*\end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \cdots \\ Z'(I-M_J)Y \end{bmatrix}$$
$$X_*'X_* \qquad \gamma \qquad = \qquad X_*'y$$

- Coeff of Determination + Multiple Corr Coef:

$$R^2 = \frac{SSReg}{Y'Y - Y'M_JY} = \frac{Y'(M-M_J)Y}{Y'Y - Y'M_JY}$$

"fraction of non-intercept SS explained by model"

- proportion of variability explained by covariates, out of total variability excl intercept.
- The sample equivalent of the mult corr coeff r^2.

- $r^2 = \max_{\alpha,\beta}\{corr(y, \alpha + \beta'x)\}$ = maximum corr with y achievable by a linear predictor.
  - Achieved by the Best Linear Predictor (BLP), so $r^2 = corr(y, \mu_y + \beta_*'(x-\mu_x))$
  - Can also write as $r^2 = \frac{\beta_*'V_{xx}\beta_*}{\sigma_y^2}$

6.3.3- Best Prediction

- Now we will treat both x + y as random variables - no observed data yet! This is purely theoretical.
- x and y have some joint distribution (x, y)
- $x = (x_1, \ldots x_{p-1})^T$, $y = (y_1, \ldots y_q)^T$ (y can be univariate or multivariate)

- Conditional Expectation: Let $m(x) = E(y|x)$. The condit. expectation $m(x)$ has the smallest MSE of any predictor of y, ie:

(*) $E[(y-m(x))^2] \le E[(y-f(x))^2]$ for univariate y. $f(x)$ is any other predictor of y.
(**) $E[(y-m(x))^T(y-m(x))] \le E[(y-f(x))^T(y-f(x))]$ for multivariate y.

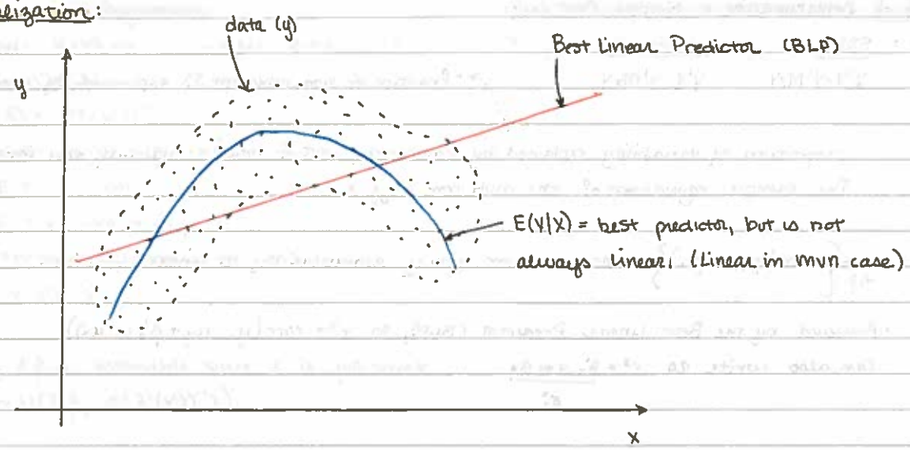
Proof (univariate case):

$$E[(y-f(x))^2] = E[(\underbrace{y-m(x)}_{a} + \underbrace{m(x)-f(x)}_{b})^2] = E[\{(y-m(x)) + (m(x)-f(x))\}\{(y-m(x)) + (m(x)-f(x))\}]$$

$$= \underbrace{E[(y-m(x))^2]}_{\text{MSE for } m(x)} + \underbrace{E[(m(x)-f(x))^2]}_{\ge 0 \text{ since squared}} + \underbrace{2E[(y-m(x))(m(x)-f(x))]}_{=0, \text{ since}}$$

$$E(y-m(x)) = E(y) - E(m(x))$$
$$= E(y) - E(E(y|x))$$
$$= E(y) - E(y)$$
$$= 0 \quad \blacksquare$$

Visualization:



4.3.4 - Best Linear Predictor (BLP)

- Any linear predictor of y has form $f(x) = \alpha + \beta' x$.
  - If y is univariate, α is scalar, β & x (p-1) length vectors.
  - If y multivariate (length q), α is (q×1), β is (p-1 × q) matrix

$$\underset{y}{\begin{bmatrix} 1 \\ y \\ 1 \end{bmatrix}_{(q+p,1)}} = \underset{\alpha}{\begin{bmatrix} 1 \\ \alpha \\ 1 \end{bmatrix}_{(q+p,1)}} + \underset{\beta^T}{\begin{bmatrix} — \beta_1 — \\ \vdots \\ — \beta_q — \end{bmatrix}_{(q\times p-1)}} \underset{(p-1)}{\begin{bmatrix} 1 \\ x \\ 1 \end{bmatrix}} = \begin{bmatrix} 1 \\ \alpha \\ 1 \end{bmatrix} + \begin{bmatrix} \beta_1' x \\ \vdots \\ \beta_q' x \end{bmatrix}$$

- $\hat{m}(x) = \hat{E}(y|x) = \mu_y + \beta_*'(x-\mu_x)$ is the <u>best</u> linear predictor (smallest MSE).
  - $\mu_y = E(y)$
  - $\mu_x = E(x)$
  - $V_{xx} = \text{Cov}(x)$
  - $V_{xy} = \text{Cov}(x,y) = V_{yx}^T$
  - $\beta_* =$ soln to $V_{xx}\beta = V_{xy}$
    ⇕
    Same as β_*' is soln to $\beta' V_{xx} = V_{yx}$
- $\hat{E}(y|x)$ is also called 'the linear expectation'.
- If $y \sim MVN$, the BLP is the best predictor, ie $E(y|x) = \hat{E}(y|x)$.

- <u>Proof that $\hat{E}(y|x)$ is best linear predictor</u>:
- Goal: Show $E[(y-\hat{E}(y|x))^2] \le E[(y-f(x))^2]$.
- Define an arbitrary linear predictor, $f(x) = \eta + \beta'(x-\mu_x)$.

$$E[(y-f(x))^2] = E\left[\left\{(y-\hat{E}(y|x)) + (\hat{E}(y|x)-f(x))\right\}^2\right]$$

$$= \underbrace{E[(y-\hat{E}(y|x))^2]}_{\text{MSE for BLP}} + \underbrace{E[(\hat{E}(y|x)-f(x))^2]}_{\ge 0 \text{ bc squared}} + \underbrace{2E[(y-\hat{E}(y|x))(\hat{E}(y|x)-f(x))]}_{\circledast = 0 \text{ as shown below}}$$

$$\circledast = E[(y-\hat{E}(y|x))(\hat{E}(y|x)-f(x))] = E\left[\left\{(y-\mu_y)-\beta_*'(x-\mu_x)\right\}\left\{\mu_y+\beta_*'(x-\mu_x)-\eta-\beta'(x-\mu_x)\right\}\right]$$

$$= E\left[\left\{(y-\mu_y)-\beta_*'(x-\mu_x)\right\}\left\{(\mu_y-\eta)+(\beta_*-\beta)'(x-\mu_x)\right\}\right]$$

$$= \underbrace{E[(y-\mu_y)(\mu_y-\eta)]}_{①} - \underbrace{E[\beta_*'(x-\mu_x)(\mu_y-\eta)]}_{②} + \underbrace{E[(\beta_*-\beta)'(x-\mu_x)(y-\mu_y)]}_{③} - \underbrace{E[\beta_*'(x-\mu_x)(x-\mu_x)'(\beta_*-\beta)]}_{④}$$

① = 0 since $E(y-\mu_y) = \mu_y - \mu_y = 0$

② = 0 since $E(x-\mu_x) = \mu_x - \mu_x = 0$ (expectation is wrt both x & y)

③ $= (\beta_*-\beta)' E((x-\mu_x)(y-\mu_y)) = (\beta_*-\beta)' V_{xy}$

④ $= \beta_*' E((x-\mu_x)(x-\mu_x)^T)(\beta_*-\beta) = \beta_*' V_{xx}(\beta_*-\beta)$

Then β_* is soln to $V_{xx}\beta = V_{xy}$, so $\beta_*' V_{xx} = V_{yx}$, so ④ $= V_{yx}(\beta_*-\beta) = (\beta_*-\beta)' V_{xy}$

Then ③ and ④ are identical, so $\circledast = 0$.

Therefore, $\hat{E}(y|x)$ has smallest MSE, so is best linear predictor. ∎

LM

Properties of Linear Expectations:

1. $\hat{E}(Ay+a|x) = A\hat{E}(y|x)+a$
2. $\hat{E}(x_i|x) = x_i$ (one of the coordinates of x)
3. $\hat{E}(x'\beta|x) = x'\beta$
4. $\hat{E}(y|Ax+b) = \hat{E}(y|x)$
5. $\hat{E}(y|x)$ if $Cov(x,y)=0$ is μ_y

Principal Components Analysis

- PCA is the problem of finding linear combos of y, $a'y$, that are Best Linear Predictors of the original y.
- Solution: eigenvecs of $Cov(y)$, ordered by largest eigval.

- **Two ways to think of PCA:**
  1) As finding BLPs for y.
  2) As finding directions a_i to project y s.t. projection onto $C(a_i)$ has smallest MSE.

- In both cases, the principal components are the first r eigenvecs of $\Sigma_y = Cov(y)$, $a_1 \dots a_r$, ordered by eigvals $\phi_1 \dots \phi_r$ s.t. $\phi_1 \geq \dots \geq \phi_r$.
- Eigvals ϕ tell what % of variance is explained by each PC.
- We center the data for PCA.

- **Projection onto $C(a_i)$ for PC a_i:** $\text{proj} = \frac{a_i^T y a_i}{\|a_i\|} = a_i^T y a$ (if $\|a_i\| \neq 1$)

- **Proof that eigvals + eigvecs are soln:**

$$\min_a \left\{ E[(y-a'ya)^T(y-a'ya)] \right\} = \min_a \left\{ E[y'y - 2y'a'ya + a'y'aa'ya] \right\} \qquad (a'a=1)$$

$$= \min_a \left\{ E[y'y - 2(a'y)^2 + (a'y)^2] \right\} = \min_a \left\{ \underbrace{E(y'y)}_{①} - \underbrace{E[(a'y)^2]}_{②} \right\}$$

① Doesn't depend on a, can ignore. $y'y = tr(y'y) = tr(yy') = tr(\Sigma_y) = \sum \phi_j$ fa Σ_y (sum of eigvals)

② $E[(a'y)(a'y)] = E(a'yy'a) = E(a'\Sigma_y a) = a'\Sigma_y a$

Then ~~minimizing~~ maximizing $tr(\Sigma_y) - a'\Sigma_y a$ is equiv to ~~min~~ max $a'\Sigma_y a$.

$\therefore \max_a \{ a'\Sigma_y a \}$ subj. to $\|a\| = a'a = 1$, otherwise could just make a very large.

- Write Lagrangian: $\ell(a) = a'\Sigma_y a - \lambda(a'a-1)$
- $\frac{\partial \ell(a)}{\partial a} = 0 \longrightarrow 2\Sigma_y a - 2\lambda a = 0 \longrightarrow \Sigma_y a = \lambda a$, so a = eigval, λ = eigvec.

✱ $Var(a'y) = a'\Sigma_y a = a'\lambda a = \lambda(a'a) = \lambda$, so eigvals explain variance. ∎

LM

Example PCA Problem: (HW 6, M2 #2)

- Let $a_1 \dots a_r$ be a set of r principal components for y.
Show that $tr\{Cov[y - \hat{E}(y|a_1'y..a_r'y)]\} = \sum_{j=r+1}^{p} \phi_j$ where ϕ_j are eigenvals.

- $Cov(y - \hat{E}(y|x)) = V_y - V_{yx} V_x^{-1} V_{xy}$ where $x = [a_1'y, \dots, a_r'y]^T_{(r\times 1)}$

- $V_y = Cov(y) = \Sigma_1$
- V_x:

$$V_x = \{Cov(a_i'y, a_i'y)\} = \{a_i'\Sigma_1 a_i\} = \begin{bmatrix} a_1'\Sigma_1 a_1 & & 0 \\ & \ddots & \\ 0 & & a_r'\Sigma_1 a_r \end{bmatrix}$$ since $a_i'a_j = 0$ for $i \neq j$ bc eigvecs orthonormal

- $V_{yx} = [Cov(y, a_1'y) \dots Cov(y, a_r'y)] = [a_1'\Sigma_1 \dots a_r'\Sigma_1]$

- Then put it together, without trace yet: $V_y - V_{yx}V_x^{-1}V_{xy}$

$$= \Sigma_1 - [a_1'\Sigma_1 \dots a_r'\Sigma_1] \begin{bmatrix} (a_1'\Sigma_1 a_1)^{-1} & & 0 \\ & \ddots & \\ 0 & & (a_r'\Sigma_1 a_r) \end{bmatrix} \begin{bmatrix} \Sigma_1 a_1 \\ \vdots \\ \Sigma_1 a_r \end{bmatrix}$$

$$= \Sigma_1 - \left(\frac{a_1'\Sigma_1\Sigma_1 a_1}{a_1'\Sigma_1 a_1} + \dots + \frac{a_r'\Sigma_1\Sigma_1 a_r}{a_r'\Sigma_1 a_r} \right)$$

Then a_i is eigval for Σ_1, so $\Sigma_1 a_i = \phi_i a_i$

$$= \Sigma_1 - \left(\frac{\phi_1^2 a_1'a_1}{\phi_1 a_1'a_1} + \dots + \frac{\phi_r^2 a_r'a_r}{\phi_r a_r'a_r} \right) = \Sigma_1 - (\phi_1 + \dots + \phi_r)$$ since $a_i'a_i = 1$ (orthonormal eigvecs)

Now add trace:

- $tr(\Sigma_1 - (\phi_1 + \dots + \phi_r)) = tr(\Sigma_1) - \sum_{j=1}^{r} \phi_j$

Then $tr(\Sigma_1)$ = sum of its eigvals; $\sum_{i=1}^{p} \phi_j$

- $= \sum_{i=1}^{p} \phi_j - \sum_{j=1}^{r} \phi_j$
- $= \sum_{j=r+1}^{p} \phi_j$ ■

Gauss-Markov Theorem

Thm: $\hat{\beta}$ is the BLUE among all linear predictors, with best = smallest variance.

Proof: Let $a'y$ be an arbitrary unbiased linear predictor.

- $E(y) = X\beta$, $Var(y) = \sigma^2 I$, $\hat{\beta} \sim N(X\beta, \sigma^2(x'x)^{-1})$. Also, $\hat{\beta} = p'My$ for $\lambda'\beta = p'X\beta$

- $Var(a'y) = Var(a'y - \hat{\beta} + \hat{\beta})$
$= Var(a'y - p'My + p'My)$
$= Var(p'My) + \underbrace{Var(a'y - p'My)}_{>0 \text{ bc is a variance}} + \underbrace{2Cov(a'y - p'My, p'My)}_{=0}$

$Cov(a'y - p'My, p'My) = (a' - p'M)\, Cov(y)\, Mp$
$= \sigma^2 (a' - p'M)Mp$
$= \sigma^2 (a'M - p'M)p = 0$

Then $E(a'y) = \lambda'\beta = p'X\beta$
And $E(a'y) = a'E(y) = a'X\beta$
} $p' = a'$, so

- Therefore, $\hat{\beta}$ is has smallest variance among all unbiased linear estimators.

Exponential Families

1. Usual parameterization: $f(x|\theta) = h(x) \cdot c(\theta) \cdot \exp\left[\sum_{i=1}^{k} w_i(\theta)\, t_i(x)\right]$
where θ is a vec with k params.

· 1-param families are just $f(x|\theta) = h(x)\, c(\theta) \exp[w_1(\theta)\, t_1(x)]$

2. Natural Parameterization: $p(x|\theta) = h(x) \cdot \exp[\theta x - \psi(\theta)]$ where $E(x) = \psi'(\theta)$.

Ex 1) Is pareto(a,b) with a fixed, b unknown an exponential family?

$$p(x|a,b) = \frac{b a^b}{x^{b+1}}, \quad a \le x \le \infty, \quad a>0, b>0$$

$$= b a^b x^{-(b+1)} \underset{[a,\infty)}{I(x)} \cdot \underset{(>0)}{I(a)} \cdot \underset{(>0)}{I(b)}$$

$$= b a^b I(x) \cdot I(a) \cdot I(b)\, e^{\log(x^{-(b+1)})} = b a^b I(x) I(a) I(b)\, e^{-(b+1)\log x}$$

Is an exp family, with $h(x) = I(x)$ $\quad w(\theta) = b+1$
$c(\theta) = b a^b I(a) I(b)$ $\quad t(x) = -\log x$

Ex 2) Is pareto(a,b) with a unknown, b fixed an exponential family?

$$p(x|a,b) = b a^b x^{-(b+1)} \cdot \underset{[a,\infty)}{I(x)} \cdot \underset{(>0)}{I(a)} \cdot \underset{(>0)}{I(b)}$$

No, bc of the $\underset{[a,\infty)}{I(x)}$ term. Cannot separate the a and x here.

Ex 3) Is $ga(a,b)$ a 2-param exponential family?

$$f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad a>0, b>0$$

$$= \frac{b^a}{\Gamma(a)} \cdot \underset{(>0)}{I(a)} \cdot \underset{(>0)}{I(b)} \cdot e^{\log\left(x^{a-1} e^{-bx}\right)}$$

$$= \frac{b^a}{\Gamma(a)} I(a)\, I(b)\, e^{(a-1)\log x + (-bx)}$$

Is a 2-param exp family, with $h(x) = 1$ $\quad w_1(\theta) = a-1$
$c(\theta) = \frac{b^a}{\Gamma(a)} I(a) I(b)$ $\quad w_2(\theta) = -b$
$t_1(x) = \log x$
$t_2(x) = x$

Ex 4) Is $X \sim \text{Binom}(n,p)$ where $n=20$ known a 1-param exp family?

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad p \in [0,1]$$

$$= \binom{n}{x} \underset{[0,1]}{I(p)}\, p^x (1-p)^n (1-p)^{-x}$$

$$= \binom{n}{x} I(p)(1-p)^n\, e^{\log(p^x (1-p)^{-x})}$$

$$= \binom{n}{x} I(p)(1-p)^n\, e^{\log\left(\frac{p}{1-p}\right)^x}$$

$$= \binom{n}{x} I(p)\,(1-p)^n\, e^{x \log\left(\frac{p}{1-p}\right)}$$

Is 1-param exp family, with $h(x) = \binom{n}{x}$ $\quad w(\theta) = \log\left(\frac{p}{1-p}\right)$
$c(\theta) = I(p)\,(1-p)^n$ $\quad t(x) = x$

Ex 5 (FY 15 Prelim #3):

· $p(x|\theta) = h(x)\exp[\theta x - \psi(\theta)]$ is natural parameterization.

A Let $\mathcal{X}$ be sample space of RV $x \sim p(x|\theta)$, from exp family w param θ. Use identities below to show $E(x) = \psi'(\theta)$.

· $\int_{\mathcal{X}} p(x|\theta)\,dx = 1$ and $\frac{\partial}{\partial\theta}\int_{\mathcal{X}} p(x|\theta)\,dx = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} p(x|\theta)\,dx$

Proof: $E(x) = \int x\,p(x|\theta)\,dx = \int x\,h(x)\exp[\theta x - \psi(\theta)]\,dx$

$= e^{-\psi(\theta)} \int x\cdot h(x)\, e^{\theta x}\,dx$ Then $x e^{\theta x} = \frac{\partial}{\partial\theta} e^{\theta x}$, so

$= e^{-\psi(\theta)} \int h(x)\cdot \frac{\partial}{\partial\theta} e^{\theta x}\,dx = e^{-\psi(\theta)} \frac{\partial}{\partial\theta}\int h(x) e^{\theta x}\,dx$ by 2nd identity.

Then $\int h(x) e^{\theta x}\,dx$ integrates to $e^{\psi(\theta)}$ bc of 1st identity.

$= e^{-\psi(\theta)} \frac{d}{d\theta}\left(e^{\psi(\theta)}\right) = \psi'(\theta)\, e^{\psi(\theta)} e^{-\psi(\theta)} = \psi'(\theta)$. ■

B Express $N(\mu,\sigma^2)$ with unknown μ, known σ^2 in natural exponential form

$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)\right]$

$= \underbrace{\frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma}}_{h(x)} \cdot \exp\left[\underbrace{\frac{\mu}{\sigma^2}}_{\theta\mu}\, x - \underbrace{\frac{\mu^2}{2\sigma^2}}_{\psi(\theta)}\right]$

C Consider a Bayesian model where $x \sim N(\theta,1)$ and $\theta \sim \pi(\theta)$. The marginal for x is

$$m(x) = \int_{\mathbb{R}} \phi(x|\theta,1)\,\pi(\theta)\,d\theta \quad \text{where } \phi \sim \text{normal density.}$$

· Prove the posterior mean can be written as $E(\theta|x) = x + \frac{\partial}{\partial x}\log m(x)$.

· Posterior: $\frac{\phi(x|\theta,1)\,\pi(\theta)}{m(x)} \cdot \frac{1}{\sqrt{2\pi}}\cdot\pi(\theta)\cdot\exp\left[-\frac{1}{2}(x-\theta)^2\right] m(x)^{-1}$

$= \frac{\pi(\theta)}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(x-\theta)^2 - \log m(x)\right]$

Now recall since working with posteria $p(\theta|x)$, x is like θ in formulas

$= \underbrace{\frac{\pi(\theta)}{\sqrt{2\pi}}\exp\left[\frac{-\theta^2}{2}\right]}_{h(\theta)} \exp\left[\underbrace{x\theta}_{x\theta} - \underbrace{\left(\frac{x^2}{2} + \log m(x)\right)}_{\psi(x)}\right]$

· From part A, know that $E(x) = \psi'(x)$, so:

$E(\theta|x) = \psi'(x) = \frac{2x}{2} + \frac{\partial}{\partial x}\log m(x) = x + \frac{\partial}{\partial x}\log m(x)$ ■

Monte Carlo Integration Basics:

- **Setting:** want to evaluate integrals of the form $I = \int g(x) f(x)\,dx$ where $f(x)$ is a density, $g(x)$ is some fctn.

- **Approach:** sample x_i's $\overset{iid}{\sim} f(x)$, then plug x_i's into $\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} g(x_i)$

- **2 Key Results:**

  1) $E(\hat{I}_N) = I$ since $E(g(x)) = \int g(x) f(x)\,dx$

  2) $\mathrm{Var}(\hat{I}_N) = \frac{1}{N^2}(\mathrm{Var}(g(x_1)) + \dots + \mathrm{Var}(g(x_N))) = \frac{N\,\mathrm{Var}(g(x))}{N^2} = \frac{\mathrm{Var}(g(x))}{N}$ since x's iid.

- **Convergence:** $\hat{I}_N \xrightarrow{P} I$ which can be shown by Chebyshev's inequality.

  - $P(|X - E(X)| \geq t \cdot \mathrm{Var}(X)) \leq \frac{1}{t^2} \equiv P(|X - E(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$

  - $P(|\hat{I}_N - E(\hat{I}_N)| \geq t) \leq \frac{\mathrm{Var}(\hat{I}_N)}{t^2}$

  $= P(|\hat{I}_N - I| \geq t) \leq \frac{\mathrm{Var}(\hat{I}_N)}{t^2}$

  $= P(|\hat{I}_N - I| \geq t) \leq \frac{\mathrm{Var}(g(x))}{nt^2}$ and $\frac{\mathrm{Var}(g(x))}{nt^2} \to 0$ as $n \to \infty$.

Therefore, $\hat{I}_N \xrightarrow{P} I$. ∎

Rejection Sampling:

- **Goal**: Sample from some density $f(x)$.

- **Steps**:

1) Write $f(x) \propto \ell(x) \cdot h(x)$ where $\cdot\ \ell(x)$ is a non-neg fctn w. upper bound M.
$\cdot\ h(x)$ is a density we can sample from.

2) Check $\ell(x)$ is bounded by taking $\frac{\partial}{\partial x}\ell(x) = 0$ and solve for root(s) x^* and $M = \ell(x^*)$.

* To get $f(x)$ in form $\ell(x) \cdot h(x)$, can split up $f(x)$, or can multiply + divide by a density.

3) Generate $U \sim U(0,1)$

4) Generate h from $h(x)$.

5) Calculate ℓ_x as $\ell(h)$.

6) Acceptance test: if $u \leq \frac{\ell_x}{M}$ then accept ℓ_x as a sample from $f(x)$. Else discard + restart.

- **Probability of acceptance**:

- If $f(x) = \ell(x) \cdot h(x)$, $p(\text{accept}) = 1/M$

- If $f(x) \propto \ell(x) \cdot h(x)$, $p(\text{accept}) \approx \frac{\#\text{accepts}}{\#\text{draws}}$ (cannot calculate directly)

- **Strategy**: Works best if can find $h(x)$ similar in shape to $f(x)$. → higher $p(\text{accept})$.

- Downside to approach: don't learn from previous samples; always starting over. ARS (adaptive rej sampling) addresses this.

Adaptive Rejection Sampling:

- **goal**: Sample from some density $f(x)$. Must be log-concave: $\frac{\partial^2 \log f(x)}{\partial^2 x} \leq 0$

- **Approach**: Construct a piecewise linear envelope from $\log f(x)$, and sample from envelope by exponentiating the linear parts (bc know how to sample from exponentials.) Done via rejection sampling, and each time a point is rejected, update the envelope to include an intersection of tangent lines at the sampled point.

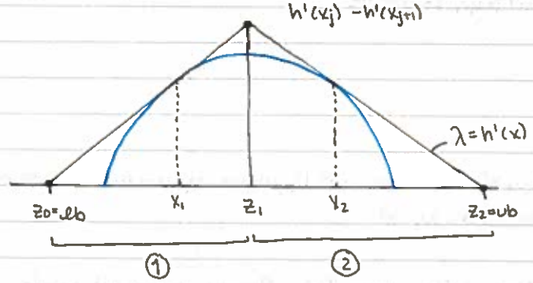
- **Steps**:

1) · Let $f(x)$ be the target fctn. Verify it is log-concave.
· Let $h(x) = \log(x)$
· Let $h'(x) = \frac{\partial}{\partial x}\log f(x)$

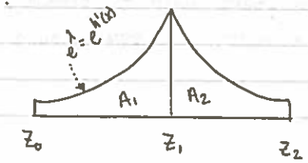
2) Begin w. a vector of at least two x vals to try. Calculate the envelope at each point, ie the slope of the tangent and the intercept which ensures that the tan line intersects the point x_i.

· slopes: $h'(x) = m$
· intercepts: $b = h(x) - x \cdot h'(x)$ (bc $y = mx + b \rightarrow h(x) = h'(x) \cdot x + b$)
· points z_i where tan lines meet, to divide envelope into segments:

$$z_j = \frac{h(x_{j+1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})}$$



3) Exponentiate the piecewise components to get a set of (rotated) truncated exponentials.



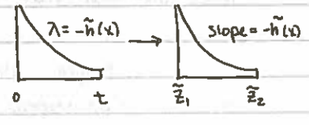
4) Calculate the area of each exponential. This is used for the first step of sampling from the exponential - must first select a segment to use, with that sampling weighted by area under each segment.

$$A_i = \frac{\exp[h(x_i) - x_i h'(x_i)]}{h'(x_i)} \left(\exp[h'(x_i) z_{i+1}] - \exp[h'(x_i) z_i] \right) \quad \text{which is} \quad \int_{z_i}^{z_{i+1}} \exp\underbrace{[h'(x_i)x + (h(x_i) - h'(x_i)x_i)]}_{\text{tan line fctn}} dx$$

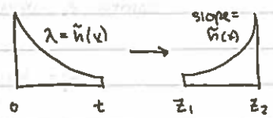
5) Sample to pick which exponential segment to use. Chooses an "area".
- selected area = sample(1:length(Areas), size=1, repl=T, prob = $\frac{\text{Areas}}{\text{sum(Areas)}}$)
- gives a selected area weighted by sample size.

6) Draw an exponential sample (size=1) from the selected exponential segment.
- Let $\tilde{h}(x)$ be the slope for the chosen segment.
- Let $\tilde{z}_1, \tilde{z}_2$ be the z's bounding the segment.
- Sample $\tilde{x}$ from the truncated exponential. Exp(λ) must have positive rate λ, so:

- If $\tilde{h}(x) < 0$, slope neg. Facing right direction, so use r=rexp(n=1, m=$-\tilde{h}(x)$, t=$z_2 - z_1$)=$\tilde{x}$ then have shifted result $x^* = z_1 + \tilde{x}$.



- If $\tilde{h}(x) > 0$, slope pos. Facing wrong direction, so use r=rexp(n=1, m=$\tilde{h}(x)$, t=$z_2 - z_1$)=$\tilde{x}$ to sample, then shift/flip, so result is $x^* = z_2 - \tilde{x}$.



7) Acceptance Test:
- Draw u from U(0,1).
- Calc proposal: $g(x^*) = \exp\{ h'(x_i)x^* + [h(x_i) - h'(x_i)x_i] \}$, value of tan line for selected segment, evaluated at x^*.

- If $u \le f(x^*)/g(x^*)$, accept x^* as being from $f(x)$. Repeat process w/ same envelope.

- If not, reject x^*. Also, update vector of x points used to create piecewise linear envelope to include x^*. Sort the x vector, and start process again with new x vector.

Importance Sampling

- Goal: Estimate the integral $I = \int g(x)\,dx$
- Approach: Write I as $I = \int \frac{g(x)}{f(x)} \cdot f(x)\,dx$ where $f(x)$ is a density we can sample.

  Then use Monte Carlo as usual to approximate the integral.

- Steps:
  1) Rewrite I as $I = \int \frac{g(x)}{f(x)} \cdot f(x)\,dx$
  2) Draw $x_i \overset{iid}{\sim} f(x)$
  3) Estimate $\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x)}{f(x)}$ where $x_i \sim f(x)$. (Even though $x_i \sim f(x)$, still plug into the whole thing, $g(x)/f(x)$.)
- Key: Now we are introducing density $f(x)$ to help us, instead of beginning with some integral $I = \int g(x)f(x)\,dx$ which already contains $f(x)$.

Normalized Importance Sampling

- Want to estimate integral $I = E_f(g(x)) = \int g(x) f(x)\,dx$, but we only know $f(x)$ up to a constant of proportionality, call this $f^*(x)$.
- Can't use regular Monte Carlo integration, bc need the constant included for integration.
- Can use Monte Carlo with a modification: $f(x) = \frac{f^*(x)}{\int f^*(x)\,dx}$ ← normalizing constant

1) Estimate normalizing constant:

$$\hat{I}^*_N = \frac{1}{N} \sum_{i=1}^{n} \frac{f^*(x)}{h(x)} \cdot h(x), \quad x_i \overset{iid}{\sim} h(x)$$

- using importance sampling, where $h(x)$ is a density we can sample.

2) Estimate whole integral:

$$\hat{I}_N = \left[\frac{\frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i) f^*(x_i)}{h(x_i)}}{\hat{I}^*_N} \right] \quad \text{using } x_i \overset{iid}{\sim} h(x).$$

Can reuse same x's; don't need to generate a new sample. Plug x's from (1) into $\frac{g(x_i) f^*(x_i)}{h(x_i)}$

Ratio of Uniforms:

- Another way to sample from a density $f(x)$.

- **Algorithm:**
  - Calculate $a = b$. $\quad a = \max_x \sqrt{f(x)}$, $\quad b = \max_x x\sqrt{f(x)}$ — $\max_x$ = roots of x which maximize $\sqrt{f(x)}$ and $x\sqrt{f(x)}$.
  - Sample $u_1 \sim \text{unif}(0, a)$, $u_2 \sim \text{unif}(0, b)$
  - Accept $x = u_2/u_1$ as being from $f(x)$ if $u_1 \le \sqrt{f(u_2/u_1)}$, else reject.

- **Why it works:**
  - Is introducing a joint density to help sample $f(x)$: $f(x,y)$ d $y \cdot \mathbb{1}(y < \sqrt{f(x)})$
    - $\int_0^{\sqrt{f(x)}} y\,dy = \frac{1}{2} f(x)$, so $\int f(x,y)\,dy = f(x)$.
  - Bivariate transformation: $u_1 = y_1$, $u_2 = yx \longrightarrow y = u_1$, $x = u_2/u_1$
    - Then $f(u_1, u_2) \propto u_1 \cdot \mathbb{1}(0 < u_1 < \sqrt{f(u_2/u_1)})|J|$ where $|J| = \begin{vmatrix} -u_2/u_1 & 1/u_1 \\ 1 & 0 \end{vmatrix} = 1/u_1$
  - To sample from the region, can sample from its bounding rectangle, then accept when (u_1, u_2) samples land in S.

- How to find the bounds of rectangle R?
  - Range for u_1 is 0 to $\max_x \sqrt{f(x)}$ → sample u_1 from $u(0, \max_x \sqrt{f(x)})$
  - For u_2, if $x \ge 0$, then $u_1 \le \max_x \sqrt{f(x)}$ and $y_1 = u_1$, $xy = u_2 \longrightarrow u_2 \le \max_x x\cdot\sqrt{f(x)}$
  - If x can take values < 0, just look at u_2^2, and you get +/- same bound:

$$\min_x x\sqrt{f(x)} \le u_2 \le \max_x x\sqrt{f(x)}$$

Basic Markov Chain Theory

- **Goal:** Generate a sample from target density $f(x)$.

- Each element in a Markov chain depends only on previous element.
- Need a **transition density** $p(x_{n+1}|x_n)$ which satisfies the **stationary condition**:
  - $f(x_{n+1}) = \int p(x_{n+1}|x_n) f(x_n)\,dx_n$
  - The stationary condition preserves $f(x)$ as the marginal density.

- Can also sample x_i's then plug into $\hat{I}_N = \frac{1}{N}\sum_{i=1}^{n} g(x_i)$ and STILL converges, since x_{i-1}'s are marginally from $f(x)$.

- **Catch:** $E(\hat{I}_N) = I$ still, but $\text{Var}(\hat{I}_N)$ increases. This is b/c $\text{Cov}(x_{n+1}, x_n) > 0$.
  - $\text{Var}(\hat{I}_N)$: $\sqrt{n}(\hat{I}_N - I) \to N(0, \sigma^2)$, $\sigma^2 = \text{Var}(g(x_i)) + 2\sum_{i=1}^{n} \text{Cov}(g(x_i), g(x_n))$

- **Reversability:** if $p(x'|x)$ satisfies $p(x'|x) f(x) = p(x|x') f(x')$ then chain is stationary.

  Proof: Let $p(x'|x)f(x) = p(x|x')f(x)$.

  Then $\int p(x'|x) f(x)\,dx = \int \underbrace{p(x|x')}_{\text{density wrt } x,\ \text{integrates to 1}} f(x')\,dx = f(x')$ so stationary cond is met.

- **Finding a transition density:**
  - want to find a transition - evaluate its covariance, $\text{Cov}(x_{n+1}, x_n)$.
  - If we have a joint density, the product of the full conditionals is the transition.
  - If don't have a joint, need to introduce one.

  - If x is in discrete space, can directly find transition matrix P.
  - If x continuous, can introduce joint via indicator fctns.

Ex 1: X in discrete space (HW 2, #2)

- The joint density $f(x,q) = q^x(1-q)^{1-x}$, $0<q<1$, $x \in \{0,1\}$. $P(x'|x) = 2\int_0^1 q^{x+x'}(1-q)^{2-x-x'}\,dq$.
- Find the stationary density.

- Since X only takes values 0 & 1, evaluate $p(x'|x)$ at each combo of x' & x. (Plug each combo into $p(x'|x)$ & eval integral.

  - $p(x'=0|x=0) = 2/3$ — $p(x'=0|x=1) = 1/3$
  - $p(x'=1|x=0) = 1/3$ — $p(x'=1|x=1) = 2/3$

  $\rightsquigarrow P = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}$

- We know $\pi^T = \pi^T P$ since want a stationary density.
- We know $\pi = [\pi_0\ \pi_1]$ where $\pi_0 + \pi_1 = 1$, since π_0 is prob for x_0, π_1 = prob for x_1.

↓

- These give us a system of 2 eqns to solve:

(1) $[\pi_1\ \pi_0]P = [\pi_1\ \pi_0]$ — The result of solving these gives $\pi_1 = \pi_0 = 1/2$, so

(2) $\pi_1 + \pi_2 = 1$ — stationary is discrete Unif(1/2) on $\{0,1\}$.

Ex 2: X in continuous space, no joint to start: (Ex 6, prob 2)

- Find a transition which has $\pi(x) = e^{-x}$, $x>0$ as the stationary.
- Need to introduce a joint $\pi(u,x)$ s.t. $p(x'|x) = \int \pi(x|u)\pi(u|x)\,dx$ where $\pi(x|u)$ & $\pi(u|x)$ are full conditionals of $\pi(u,x)$.
- Try $\pi(u,x) = \mathbb{1}(u<f(x))$, where $f(x) = e^{-x}$. Find full conditionals:

  - $\pi(u|x) = U(0, e^{-x})$
  - $\pi(x|u)$ is unif on $x: f(x) > u \rightarrow e^{-x} > u \rightarrow -x > \log u \rightarrow x < -\log u$
    so $\pi(x|u) = U(0, -\log u)$.

- Begin w/ any x_1, then iterate sampling u from $U(0, e^{-x})$
  x from $U(0, -\log u)$
- Can also find the stationary which gives update $x' = \phi(x,u,v)$ where $u,v \overset{iid}{\sim} U(0,1)$.

  - $u \sim U(0, e^{-x}) \rightarrow u = e^{-x}v$ (just rescaling $v \sim U(0,1)$).
  - $x \sim U(0, -\log u) \rightarrow U(0, -\log(e^{-x}v)) = U(0, x - \log v)$

  ↓

  $x' = u\cdot(x_n - \log v)$ with $u,v \overset{iid}{\sim} U(0,1)$

*why? ease of sampling, & can make finding $cov(x',x)$.

Covariance of Markov Chains

- For a stationary density, first find the transition, then find $Cov(x_{n+1}, x_n)$ i.e. $Cov(x', x)$.
- Two helpful tricks for finding covariance:
  1) Find transition in terms of $x' = \phi(x,u,v)$ where $u,v \overset{iid}{\sim} U(0,1)$
  2) Iterated expectation: $Cov(x',x) = E(x'x) - E(x')E(x) = E\left(x\cdot E(x'|x)\right) - E(x)\cdot E\left(E(x'|x)\right)$

- **Ex: Exercise 6, Prob 1:**
- Find a transition which has $\pi(x) \sim Ga(x|a,1)$ as stationary, s.t. $\pi(x) = \frac{x^{a-1}e^{-x}}{\Gamma(a)}$, $x>0, a>$
  Find $cov(x',x)$.

1) Introduce joint density & find transition:

- Need a joint $\pi(u,x)$ so can use full conditionals $\pi(u|x)$, $\pi(x|u)$.
- $\mathbb{1}\left(u < \frac{x^{a-1}e^{-x}}{\Gamma(a)}\right)$ not easily invertible. Split into product of 2 fcns & take intersection.
- $\pi(u,x) = \underbrace{\mathbb{1}\left(u < \frac{x^{a-1}}{\Gamma(a)}\right)}_{\pi_1(u,x)} \cdot \underbrace{\mathbb{1}\left(u < e^{-x}\right)}_{\pi_2(u,x)}$ Then need two pairs of full conditionals.

a) For $\pi_1(u,x)$: $\pi_1(u|x) = U\left(0, \frac{x^{a-1}}{\Gamma(a)}\right) \rightarrow u < \frac{x^{a-1}}{\Gamma(a)} \rightarrow \log(u\Gamma(a)) < (a-1)\log x$

$\rightarrow \exp\left[\frac{\log(u\Gamma(a))}{a-1}\right] < x$

$\pi_1(x|u) = U\left(\exp\left[\frac{\log u\Gamma(a)}{a-1}\right], \infty\right)$

b) For $\pi_2(u,x)$ $\pi_2(u|x) = U(0, e^{-x}) \rightarrow u < e^{-x} \rightarrow x < -\log u$
$\pi_2(x|u) = U(0, -\log x)$

- Can sample $u_1, u_2 \overset{iid}{\sim} U(0,1)$ for the two bounds, then combine the above to get

$$x' \sim U\left(\exp\left[\frac{\log u_1\Gamma(a)}{a-1}\right], -\log u_2\right)$$

2) Put above result in form $x' = \phi(x,u,v)$ for easy covariance calcs. In this case, we will need a 3rd uniform:

$x' = \phi(x,u,y,w)$ where $u,y,w \overset{iid}{\sim} U(0,1)$

· $\pi_1(u_1|x) \longrightarrow u_1 = v\,\frac{x^{a-1}}{\Gamma(a)}$

· $\pi_2(u_2|x) \longrightarrow u_2 = we^{-x}$

} for $v, w \overset{iid}{\sim} U(0,1)$

· Plug u_1 & u_2 into x bounds: $\pi(x|u_1,u_2) \sim U\left(\exp\left[\frac{\log u_1\Gamma(a)}{a-1}\right], -\log u_2\right) \rightarrow U\left[xe^{\log v/a-1}, x-\log w\right]$

· Add $Y \sim U(0,1)$. when $Y=0$, want lower bound for x', ie $xe^{\log v/a-1}$.
when $Y=1$, want upper bound for x', ie $x - \log w$.

$$x' = x\,e^{\log v/a-1} + Y\left[(x-\log w) - x\,e^{\log v/a-1}\right] \quad \text{where } u, w, Y \overset{iid}{\sim} U(0,1)$$

· $xe^{\log v/a-1} = x\,v^{1/a-1}$

3) Find $Cov(x', x)$.

· $Cov(x', x) = E(x'x) - E(x)E(x') = E\left(x\,\underbrace{E(x'|x)}_{\bullet}\right) - E(x)\cdot\underbrace{E(E(x'|x))}_{\bullet}$

• $E(x'|x) = E\left(xv^{1/a-1} + yx - y\log w - yxv^{1/a-1} \mid x\right)$ by plugging in x' expression

$= xE(v^{1/a-1}) + xE(y) - E(y)E(\log w) - xE(y)E(v^{1/a-1})$ and recall $y, v, w \overset{iid}{\sim} u(0,1)$

$\int_0^1 v^{1/a-1}(1)\,dv = \frac{a-1}{a}$ $\quad = 1/2$ $\quad = 1/2$ $\quad \int_0^1 \log(w)(1)\,dw = -1$ $\quad = \frac{a-1}{a}$

$= x\left(\frac{a-1}{a}\right) + \frac{x}{2} - \frac{1}{2}(-1) - x\left(\frac{1}{2}\right)\left(\frac{a-1}{a}\right) = \boxed{\frac{1}{2}x\left(\frac{a-1}{a}+1\right)+\frac{1}{2}}$ Then plug back in:

$$Cov(x', x) = E\left[x\left(\tfrac{1}{2}x\left(\tfrac{a-1}{a}+1\right)+\tfrac{1}{2}\right)\right] - E\left[\tfrac{1}{2}x\left(\tfrac{a-1}{a}+1\right)+\tfrac{1}{2}\right]E(x)$$

$$= E\left[\frac{1}{2}x^2\left(\frac{a-1}{a}+1\right) + \frac{x}{2}\right] - \left[\frac{1}{2}x\left(\frac{a-1}{a}+1\right)+\frac{1}{2}\right]E(x)$$

$$= \frac{a-1}{2a}E(x^2) + \frac{1}{2}E(x^2) + \frac{1}{2}E(x) - \frac{a-1}{2a}E(x)^2 - \frac{1}{2}E(x)^2 - \frac{1}{2}E(x)$$

$$= \left(\frac{1}{2} - \frac{1}{2a}\right)Var(x) + \frac{1}{2}Var(x)$$

$$= \left(1 - \frac{1}{2a}\right)Var(x)$$

· what we really want is AR1, ie autocorrelation:

$$\rho = \frac{Cov(x', x)}{sd(x')\,sd(x)} = \frac{\left(1-\frac{1}{2a}\right)Var(x)}{[sd(x)]^2 = Var(x)} = \boxed{\left(1 - \frac{1}{2a}\right)}$$

since $x' \perp$ iid

Metropolis-Hastings:

· goal: Sample a chain from density $\pi(x)$.

· Algorithm:
· Draw x^* from proposal $q(x'|x)$.
· Draw u from $U(0,1)$.
· if $u < \alpha(x^*|x)$, then $x' = x^*$, else $x' = x$.

$$\cdot\ \alpha(x'|x) = \min\left\{\frac{\pi(x')}{\pi(x)}\cdot\frac{q(x|x')}{q(x'|x)},\ 1\right\}$$

$$\cdot\ p(x'|x) = \alpha(x^*|x)\,q(x^*|x) + (1-r(x))\,\mathbb{1}(x^*=x)\ ;\ r(x) = \int \alpha(x^*|x)\,q(x^*|x)\,dx^*$$

· Proof: Show the stationary condition holds for target $\pi(x)$. Is sufficient to show reversability, ie $p(x'|x)\pi(x') = p(x|x')\pi(x)$

$$p(x'|x)\pi(x) = \pi(x)\cdot\left[\alpha(x^*|x)\,q(x^*|x) + \underbrace{(1-r(x))\,\mathbb{1}(x'=x)}\right]$$

$= (1-r(x'))\,\mathbb{1}(x=x')$ since only nonzero when $x=x'$

$$= \pi(x)\left[\min\left\{1, \frac{\pi(x')}{\pi(x)}\cdot\frac{q(x|x')}{q(x'|x)}\right\}q(x'|x) + (1-r(x'))\,\mathbb{1}(x=x')\right]$$

Multiply $\pi(x)$ through, & multiply $q(x'|x)$ into the min fctn.

$$= \min\left\{\pi(x)\,q(x'|x),\ \frac{\pi(x')\pi(x)}{\pi(x)}\,\frac{q(x|x')\,q(x'|x)}{q(x'|x)}\right\} + \underbrace{\pi(x)}_{=\pi(x')\text{ here, since}\neq 0\text{ when }x=x'.}(1-r(x'))\,\mathbb{1}(x'=x)$$

$$= \pi(x')\left[\min\left\{\frac{\pi(x)\,q(x'|x)}{\pi(x')\,q(x'|x)},\ 1\right\}\overset{q(x|x')}{*}(1-r(x'))\,\mathbb{1}(x'=x)\right]$$

by pulling $\pi(x')$ out and pulling $q(x|x')$ out of min fctn

· $\pi(x')\,p(x|x')$. Therefore, reversability holds, so stationary holds. ∎

· How to pick proposal q?

· Want density w. mass in similar areas as $\pi(x)$, w. same support.

· Center q around previous x value.

· If variance of q is too small, chain will get stuck around same x value, ie high acceptance, slow exploration of the space.

· If variance of q is too ~~small~~ large, chain will bounce around, ie fast exploration of space, low acceptance.

· Plot the traces of chains to evaluate how proposal choice is mixing.

· Example: (Midterm, #3)

· Want to sample from the power law dist, $p(X=x) \propto x^{-\alpha}$, for $x \in \{1,2,3\ldots\}$, $\alpha > 1$.

Will use MH with proposal $q(x'|x) = \begin{cases} x' = x+1 \text{ with prob } 1/2 \\ x' = x-1 \text{ with prob } 1/2 \end{cases}$ and $q(2|1) = 1$.

Write the details of implementing MH.

· First, define $\alpha(x',x)$:

· For $x>1$, $\alpha(x',x) = \min\left\{1, \frac{f(x)q(x|x')}{f(x')q(x'|x)}\right\} = \min\left\{\frac{x^{-\alpha}(1/2)}{(x')^{-\alpha}(1/2)}, 1\right\} = \min\left\{\frac{x^{-\alpha}}{(x')^{-\alpha}}, 1\right\}$

· For $x=1$, $\alpha(x',x) = \min\left\{\frac{x^{-\alpha}(0)}{(x)^{-\alpha}(1)}, 1\right\} = 0$

· Algorithm:

1. Draw x^* from proposal.
   · If $x>1$, $x^* = \text{Binom}(1, 1/2)$ outcome for $x^* = x+1$ or $x^* = x-1$.
   · If $x=1$, $x^* = 2$.
2. Draw $u \sim U(0,1)$
3. Calc $\alpha(x^*, x)$ based on above definition.
4. If $u < \alpha(x^*, x)$ then $x' = x^*$, else $x' = x$.

Note: Since $\alpha = 0$ if $x=1$, we will always accept $x' = x^*$ for the move to 2. ■

Gibbs Sampler

· MH is difficult for hi-dim proposals. The Gibbs sampler is an alternative. Great for generating samples from a joint posteria, or any joint density.

· If $p(\mu,\lambda|x)$ is a joint density.
   · Begin with some arbitrary μ, λ.
   · Sample μ' from full conditional $\pi(\mu|\lambda,x)$
   · Sample λ' from full conditional $\pi(\lambda|\mu',x)$
   · Continue iterating. (Order doesn't matter)

· Key: product of the iteratively sampled full conditionals is the transition density:

$$p(\mu',\lambda'|x,\mu,\lambda) = \pi(\mu'|\lambda',x) \cdot \pi(\lambda'|\mu,x)$$

Proof: Prove the stationary condition holds for Gibbs. Say stationary density is $p(\mu,\lambda|x)$. Transition is $p(\mu'|\lambda',x)\,p(\lambda'|\mu,x)$.

$$\iint p(\mu'|\lambda',x)\,p(\lambda'|\mu,x)\,p(\mu,\lambda|x)\,d\mu d\lambda$$

$$= \int p(\lambda'|\mu,x)\,p(\mu,\lambda|x)\,d\mu = \int \underbrace{p(\lambda',\mu,\lambda|x)}_{\text{density wrt } \mu} d\mu = p(\lambda',\lambda|x) \quad \text{Then plug in.}$$

$$= \int p(\mu'|\lambda',x)\,p(\lambda',\lambda|x)\,d\lambda = \int \underbrace{p(\mu',\lambda',\lambda|x)}_{\text{density wrt } \lambda} d\lambda = p(\mu',\lambda'|x). \quad ■$$

* Notes:

1. In univariate case, we were introducing "u" placeholders to create a joint. No need here.
2. The $|x$ can be removed and all results would still hold.

Gibbs Sampler With MH Step:

· If one (or more) of the full conditionals is difficult to sample from, can introduce a MH step for that full conditional within the Gibbs sampler.

Gibbs Sampler Proof for 3 variables.

- Stationary is $p(\mu, \lambda, \theta | x)$
- Update order: λ, then θ, then μ. (Doesn't matter, can do in any order.)

- Transition: $p(\mu' | \theta', \lambda', x) \cdot p(\theta' | \lambda', \mu, x) \cdot p(\lambda' | \mu, \theta, x)$

- Proof that the stationary condition holds: (Begin w/ RHS)

$$\iiint p(\mu' | \theta', \lambda', x)\, p(\theta' | \lambda', \mu, x)\, \underbrace{p(\lambda' | \mu, \theta, x)\, p(\mu, \theta, \lambda | x)}\, d\lambda\, d\theta\, d\mu$$

$$= \int \underbrace{p(\lambda', \mu, \theta, \lambda | x)}_{\text{density wrt } \lambda}\, d\lambda = p(\lambda', \mu, \theta | x). \quad \text{Plug back in:}$$

$$= \iint p(\mu' | \theta', \lambda', x)\, \underbrace{p(\theta' | \lambda', \mu, x)\, p(\lambda', \mu, \theta | x)}\, d\theta\, d\mu$$

$$= \int \underbrace{p(\theta', \lambda', \mu, \theta | x)}_{\text{density wrt } \theta} d\theta = p(\theta', \lambda', \mu | x). \quad \text{Plug back in:}$$

$$= \int p(\mu' | \theta', \lambda', x)\, p(\theta', \lambda', \mu | x)\, d\mu$$

$$= \int \underbrace{p(\mu', \theta', \lambda', \mu | x)}_{\text{density wrt } \mu} d\mu = p(\mu', \theta', \lambda' | x). \quad \text{Stationary condition holds.} \blacksquare$$

2-Component Gaussian Mixture Model:

- Model: $g(x|\theta) = \omega\, N(x | \mu_1, \sigma_1^2) + (1-\omega) N(x | \mu_2, \sigma_2^2)$, $\theta = \{\omega, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$

- Problem: the lhood is difficult to use directly bc it has a sum:

$$L(\theta | x) = \prod_{i=1}^{n} \left[\omega\, N(x_i | \mu_1, \sigma_1^2) + (1-\omega) N(x_i | \mu_2, \sigma_2^2) \right]$$

- Solution: Introduce one or more latent variables to get rid of the sum in the lhood. Only need 1 latent var in this case.

$d_i = \{1, 2\}$ to tell us which component ea. obs x_i belongs to.

- Can then rewrite $g(x|\theta)$ as $g(x, d | \theta) = \omega_d N(x | \mu_d, \sigma_d^2)$ where $\omega_d \Rightarrow \omega_1 = \omega$, $\omega_2 = (1-\omega)$.

- Lhood becomes $\prod_{d_i=1} N(x_i | \mu_1, \sigma_1^2) \cdot \prod_{d_i=2} N(x_i | \mu_2, \sigma_2^2)$, ie $\prod_{i=1}^{n} \omega_{d_i} N(x_i | \mu_i, \sigma_i^2)$

- Key: Choose latent so when you integrate it out, you get original lhood.

General M-Component Gaussian Mixture Model:

- Model: $g(x|\theta) = \sum_{j=1}^{M} \omega_j N(x | \mu_j, \sigma^2)$ where $\sum \omega_j = 1$. $\theta = \{\omega_1 \ldots \omega_m, \mu_1 \ldots \mu_m, \sigma^2\}$

- Priors of μ_j's are iid Normals $N(m, \tau^2)$
- Prior for $\sigma^2 = IG(a, b)$ (Jeffreys, $1/\sigma^2$, would work also)
- Prior for $(\omega_1 \ldots \omega_m) \sim$ Dirichlet:

  - $p(\omega) \propto \prod_{j=1}^{M} \omega_j^{d_j - 1}$ with $\sum_{j=1}^{m} \omega_j = 1 \longrightarrow p(\omega) \propto \prod_{j=1}^{m-1} \omega_j^{d_j - 1} (1 - \omega_1 - \ldots - \omega_{m-1})^{d_j}$

  - If $M=2$, this is Beta. Can use Beta(1,1).

- Again introduce latent d_i's to identify class membership for each obs.

⊛

Mixture Model Example (Ex 11.4.1)

- Model: $g(x|\theta) = w\,N(x|\mu_1,\sigma^2) + (1-w)N(x|\mu_2,\sigma^2)$.

- $p(\sigma^2) \sim IG(1,1)$
- $p(\mu_1), p(\mu_2) \overset{iid}{\sim} N(0,v)$ $(m=0)$
- $p(w) \sim Beta(1,1)$.

- Use gibbs Sampler to generate samples from $\theta = \{\mu_1, \mu_2, \sigma^2, w\}$. Note: latent vars must be included in the sampler.

- Let $d = \{1,2\}$ with probs $p(d=1)=w$ and $p(d=2)=(1-w)$.

- Lhood: $\prod_{i=1}^{n} w_{di}\, N(x_i|\mu_i,\sigma^2) = \underbrace{\prod_{i=1}^{n_1} N(x_i|\mu_1,\sigma^2)}_{\text{for } d_i=1 \text{ obs}} \cdot \underbrace{\prod_{i=1}^{n_2} N(x_i|\mu_2,\sigma^2)}_{\text{for } d_i=2 \text{ obs}}$

- Full conditionals:

1) $(\mu_1|\ldots) \propto \exp\left[\frac{-1}{2v}\mu^2\right] \cdot \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2\right]$ using only obs in d_1 group

$$\propto \exp\left[\frac{-1}{2v}\mu^2 - \frac{1}{2\sigma^2}\left(n_1\mu_1^2 - 2\mu_1\sum_{i=1}^{n_1}x_i\right)\right]$$

$$\propto \exp\left[\frac{-1}{2v}\mu^2 - \frac{n_1}{2\sigma^2}\left(\mu_1^2 - 2\mu_1\underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1}x_i}_{\bar{x}_1}\right)\right]$$

$\sim$ Normal with $\text{var} = \left[\frac{1}{v} + n_1/\sigma^2\right]^{-1}$, $\text{mean} = \text{var}\cdot\left[\frac{n_1}{\sigma^2}\bar{x}_1\right]$

2) $(\mu_2|\ldots)$ is same as μ_1, but using only obs in d_2 group, and μ_2.

3) $(w|\ldots) \propto \underbrace{\prod_{i=1}^{n_1} w \cdot \prod_{i=1}^{n_2}(1-w)}_{\text{lhood}} \cdot \underbrace{w^{(1-1)}(1-w)^{(1-1)}}_{\text{Beta}(1,1)\text{ prior}} = w^{n_1}(1-w)^{n_2} \sim Beta(n_1+1, n_2+1)$

4) $(\sigma^2|\ldots) \propto (\sigma^2)^{-n/2} \underbrace{\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2\right]}_{\text{for } d_i=1 \text{ obs}} \underbrace{\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n_2}(x_i-\mu_2)^2\right]}_{\text{for } d_i=2 \text{ obs}} (\sigma^2)^{-a-1}\exp[-b/\sigma^2]$

$$\sim IG\left(\frac{n}{2}+a,\; b + \frac{1}{2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \frac{1}{2}\sum_{i=1}^{n_2}(x_i-\mu_2)^2\right)$$

5) $p(d_i=1) = \dfrac{w\cdot N(x_i|\mu_1,\sigma^2)}{w\cdot N(x_i|\mu_1,\sigma^2) + (1-w)N(x_i|\mu_2,\sigma^2)}$

Modeling 1

Convergence in Probability:

- An estimator $\hat{\theta}$ is consistent if $\hat{\theta} \xrightarrow{p} \theta$. ⊛ $\hat{\theta} \xrightarrow{p} \theta \rightarrow E(\hat{\theta}) \xrightarrow{p} E(\theta)$

- Definition: $X_n \xrightarrow{p} X$ iff $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$

- Tools:

1. Markov's Inequality: $P(g(X) \geq r) \leq \frac{E(g(X))}{r}$

2. Chebyshev's Inequality: $P(|X - E(X)| \geq t\,Var(X)) \leq \frac{1}{t^2}$ ie $P(|X - E(X)| \geq t) \leq \frac{Var(X)}{t^2}$

   (X is some parameter, ie $\hat{\theta}$, S, $\bar{X}$, etc)

3. WLLN: $\bar{X} \xrightarrow{p} \mu$. Also includes moments: $\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{p} \mu_k$

4. Continuous Mapping Theorem: If $g(x)$ continuous and $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$

   - Also works w. 2+ variables: Ex, if $g(a,b) = a/b$ and $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, then $X_n/Y_n \xrightarrow{p} X/Y$.

Ex: Generate $(Z_1, X_1)\ldots(Z_n, X_n)$ as $Z_i = \begin{cases} \text{blue w. prob } \theta \\ \text{red w. prob } 1-\theta \end{cases}$ $X_i = \begin{cases} \text{Pois}(1) \text{ if } Z_i \text{ blue} \\ \text{Pois}(4) \text{ if } Z_i \text{ red} \end{cases}$

Given only the X_is, find a consistent estimator of θ.

$E(X_i) = E(E(X_i|Z_i)) = E(1\theta + 4(1-\theta)) = 4 - 3\theta \longrightarrow E(X_i) = 4 - 3\hat{\theta} \rightarrow \hat{\theta} = \frac{4 - \bar{X}}{3}$

Consistent b/c by WLLN, $\bar{X} \xrightarrow{p} E(X_i)$, and:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = \lim_{n\to\infty} P\left(\left|\frac{4-\bar{X}}{3} - \frac{4-E(X)}{3}\right| \geq \varepsilon\right) = \lim_{n\to\infty} P(|\bar{X} - E(X)| \geq \varepsilon) = 0$$

by WLLN + continuous mapping thm.

- Consistency ≠ Unbiased:
- consistency means $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$. True if sampling dist of $\hat{\theta}$'s var $\to 0$, it becomes ↑ concentrated as $n \to \infty$.
- unbiased applies to any sample size. 'On avg, $\hat{\theta}$ hits true θ'.
- $\hat{\sigma}^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ is biased but consistent.

Convergence in Distribution:

- Definition: $X_n \xrightarrow{d} X$ if $\lim_{n\to\infty} F_n(t) = F(t)$, ie if $\lim_{n\to\infty} P(X_n \leq t) = P(X \leq t)$

- Tools:

1. Convergence in prob → convergence in distr. (Converse false)

2. Slutsky's Lemma: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, a constant, then
   - $X_n Y_n \xrightarrow{d} Xc$
   - $X_n + Y_n \xrightarrow{d} X + c$

3. CLT: If $X_1 \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\bar{X} \xrightarrow{d} N(\mu, \sigma^2/n)$

4. Continuous Mapping Thm: If $g(x)$ continuous and $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

5. Delta Method: If $\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$ and $g(x)$ is a fctn where $g'(x) \neq 0$, then

   $g(X_n) \xrightarrow{d} N(g(\mu), \frac{\sigma^2}{n} g'(\mu)^2)$

- How to begin: $P(X_n \leq t)$ and begin rearranging.

- Ex: Let $X_1 \ldots X_n \sim U(\theta, 1)$. Show that $n(\hat{\theta} - \theta) \xrightarrow{d} \text{Exp}(\lambda)$, where $\hat{\theta}$ is MLE of θ. Find λ.

- MLE: $\prod_{i=1}^{n} \frac{1}{1-\theta} = \left(\frac{1}{1-\theta}\right)^n$ is maximized for sample min, $X_{(1)}$. So $\hat{\theta}_{MLE} = X_{(1)}$.

- $P(n(\hat{\theta} - \theta) \leq t) = P(\hat{\theta} - \theta \leq t/n) = P(\min(X_i) \leq \theta + t/n) = 1 - P(\text{all } X_1 \ldots X_n \geq \theta + t/n)$

  $= 1 - P(X_1 \geq \theta + t/n) + \ldots + P(X_n \geq \theta + t/n)$ since x's iid $= 1 - P(X \geq \theta + t/n)^n$ since x's iid

  $= 1 - [1 - P(X \leq \theta + t/n)]^n$ where $X \sim U(0,1)$ so cdf is $\frac{x-a}{b-a}$, $x = \theta + t/n$, $a = \theta$, $b = 1$.

  $= 1 - \left[1 - \frac{\theta + t/n - \theta}{1 - \theta}\right]^n = 1 - \left[1 - \frac{t}{n(1-\theta)}\right]^n$ and $\lim_{x\to\infty}\left[1 - \frac{a}{x}\right]^{bx} = e^{ab}$, with $x = n$, $b = 1$, $a = \frac{-t}{1-\theta}$,

  so $= 1 - \exp\left[\frac{-t}{1-\theta}\right]$ which is the cdf for $\exp(\lambda)$ where $\lambda = \frac{1}{1-\theta}$ ■

MLE Review:

- **To find MLE:**
  - Lhood = $\prod_{i=1}^{n} f(x_i)$
  - Take log. Take $\frac{d}{\partial x} \log_e$, set $=0$, solve for param(s) $\hat{\theta}$.

- **MLE Properties:**
  1) **Consistency:** $\hat{\theta}_{mle} \xrightarrow{p} \theta$
  2) **Asymptotically normal.** (EXCEPT: 1. if θ is in support of distribution
     2. if # params grows as n grows (Neyman Scott))
  3) **Invariance:** If $\hat{\theta}$ is mle of θ, $g(\hat{\theta})$ is mle of $g(\theta)$.
  4) **Asymptotic Optimality:** Smallest variance among all unbiased estimators.

  - Asymptotic Rel. Efficiency: ratio of variances of 2 estimators.
    smaller variance = ↑ efficient.

- **Fisher Info:**
  - Asymptotic distribution of $\hat{\theta}_{mle}$ is $N \sim (\theta, I^{-1}(\theta))$ where $I(\theta)$ = Fisher's Info.
  - $I(\theta) = -E\left[\frac{\partial^2 \log f(x_i|\theta)}{\partial^2 \theta}\right] = -E\left[\frac{\partial^2 \log L}{\partial^2 \theta}\right]$

- Ex: $X_1 \dots X_n \sim$ Bernoulli(p). $L = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$

$$\log L = \Sigma x_i \log p + (n - \Sigma x_i)\log(1-p)$$

$$\frac{d}{dp}\log L = \frac{\Sigma x_i}{p} - \frac{(n-\Sigma x_i)}{1-p} \longrightarrow \frac{\partial^2 \log L}{\partial^2 p} = -\frac{\Sigma_i x_i}{p^2} - \frac{n-\Sigma x_i}{(1-p)^2} = -\frac{n\bar{x}}{p^2} - \frac{n-n\bar{x}}{(1-p)^2}$$

$$-E_p\left[\frac{\partial^2 \log L}{\partial^2 p}\right] = -E\left[\frac{-n\bar{x}}{p^2} - \frac{n-n\bar{x}}{(1-p)^2}\right] = \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p(1-p)} = I_n(p)$$

- Matrix form: elements are $\{I(\theta)\}_{i,j} = -E\left[\frac{\partial^2 \log L}{\partial^2 \theta_i \partial^2 \theta_j}\right]$

- **Jeffreys Prior:** $p(\theta) = \sqrt{\det I(\theta)}$

Cramer-Rao Lower Bound: Formalizes property that MLE has smallest var of all unbiased estimators.

$$\text{Var}(\hat{\theta}) \geq I(\theta)^{-1} \qquad \text{· Recall, } \text{Var}(\hat{\theta}_{mle}) = I(\theta)^{-1}$$

When MLEs Don't Exist:

1) Sometimes when θ is in support. (May exist still, but not asymp. normal.
   Ex: $X \sim U(0,\theta)$. $\hat{\theta} = X_{max}$ is not in interval, so MLE doesn't exist. Would if $U(0,\theta]$.
2) When p grows as n grows (Neyman Scott Example.)

- Consider n groups of normals, w. different means + shared variance.
  - $x_{11} \dots x_{1K} \sim N(\mu_1, \sigma^2)$ ... $x_{n1} \dots x_{nK} \sim N(\mu_K, \sigma^2)$

- Goal: estimate σ^2. Seems smart to take mean of sample vars from ea group.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2 \quad \text{where } s_i^2 = \frac{1}{K}\sum_{j=1}^{K}(x_{ij} - \bar{x}_j)^2$$

- We know $\frac{(k-1)s_i^2}{\sigma^2} \sim \chi^2_{k-1}$ so has mean = df = k-1.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(k-1)s_i^2}{k} = \frac{\sigma^2}{kn}\sum_{i=1}^{n}\frac{(k-1)s_i^2}{\sigma^2} \xrightarrow{WLLN} \frac{\sigma^2}{k}E\left[\frac{(k-1)s_i^2}{\sigma^2}\right] = \frac{\sigma^2\sigma^2(k-1)}{\sigma^2 k} = \frac{\sigma^2(k-1)}{k} \neq \sigma^2 \text{ as } n \to \infty$$

Bootstrap

1. Nonparametric bootstrap:

· Let $x_1 \dots x_n \sim f(x|\theta)$ where θ unknown. To estimate θ:

For b in 1:B{
- x_b^* = Sample of size n, with repl, from $x_1 \dots x_n$
- $\hat{\theta}_b$ = mle of θ, calculated using x_b^*.

}

· Then have $\hat{\theta}_1 \dots \hat{\theta}_B$. Calc $\hat{\theta} = \text{mean}(\hat{\theta}_1 \dots \hat{\theta}_B)$

$\text{Var}(\hat{\theta}) = \text{var}(\hat{\theta}_1 \dots \hat{\theta}_B)$

· Obtain CI using quantiles of $\hat{\theta}_1 \dots \hat{\theta}_B$, or $\hat{\theta} \pm 1.96\sqrt{\text{var}(\hat{\theta})}$

2. Parametric Bootstrap:

Similar to non-parametric, except:

· Begin by calcing $\hat{\theta}_{mle}$ from orig sample $x_1 \dots x_n$.

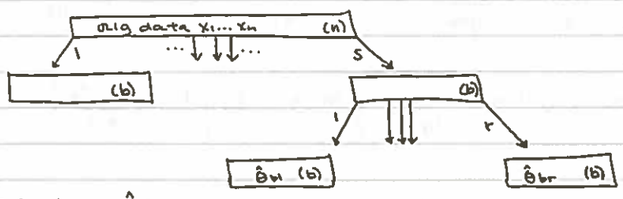
· For each bootstrap iteration, instead of drawing sample from $x_1 \dots x_n$, generate n random obs from $f(x|\hat{\theta}_{mle})$.

3. Little Bag of Bootstraps:

· Way to parallelize when data set too large for a single bootstrap.

· Sample without replacement the orig. data, into s samples of size b.

· For each sample:
  - Resample & compute $\hat{\theta}$, until total sample size = n
  - Take average of $\hat{\theta}_1 \dots \hat{\theta}_s$
  - Careful of scaling: Ex: $\hat{var} = \frac{\sum(x_i - \bar{x})^2}{n}$, but we, $\frac{\sum(x_i - \bar{x})^2}{b}$, so mult avg by (b/n)



· (b) = sample of size b

· All samples are w.o. repl.

until b·r = n

· Estimating $\hat{\text{Var}}\theta$:

Have $\text{Var}_1 = \text{Var}(\theta_{11} \dots \theta_{1b})$ to $\text{Var}_s = \text{Var}(\theta_{s1} \dots \theta_{sb})$. $\hat{\text{Var}} = \text{mean}(\text{Var}_1 \dots \text{Var}_s)$

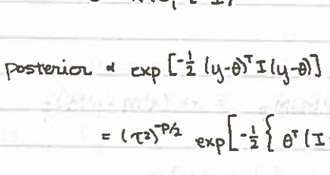
· Scaling: $\text{Var}_1 = \frac{1}{b}\sum_{i=1}^{b}(\theta_{1i} - \bar{\theta}_1)^2$, which is why need $\text{Var}_1 = \text{Var}_1\left(\frac{b}{n}\right)$ to ensure right scale.

Shrinkage Estimators

1. Dominance: Consider $\hat{\theta}$, an estimate of θ. This $\hat{\theta}$ dominates another estimator if $\hat{\theta}$ has smaller or equal MSE for all values $\theta \in \Theta$.

· $\text{MSE}(\hat{\theta}) \leq \text{MSE}(\tilde{\theta})$, ie $E[(\hat{\theta}-\theta)^2] \leq E[(\tilde{\theta}-\theta)^2]$

2. Admissible: $\hat{\theta}$ is admissible if no other estimator dominates it.



· $\tilde{\theta}$ dominates both $\hat{\theta}$ and θ^*. If these are the only 3 estimators, $\tilde{\theta}$ is admissible.

· $\hat{\theta}, \theta^*$ do not dominate ea other - neither has lower MSE over the whole param space of θ.

· To prove: Let $\tilde{\theta}$ be some arbitrary estimator. Show $E[(\hat{\theta}-\theta)^2] \leq E[(\tilde{\theta}-\theta)^2] \ \forall\theta$.

3. James Stein Estimator

· Let $y \sim N(\mu, \sigma^2 I)$. Want to estimate μ for a single obs.

· If n=1, $\hat{\mu} = \bar{y} = y_i$; just the value of that point.

· $\hat{\mu}_{JS} = \left[1 - \frac{(p-2)}{\|y_i\|^2}\right] y_i$

· Dominates $\hat{\mu}_{mle}$ for $p \geq 3$

· Adds a bit of bias in exchange for lower variance; shrinks est. of μ towards 0.

4. Empirical Bayes Estimator:

· Let $y|\theta \sim N(\theta, I)$ Do normal-normal update:

$\theta \sim N(0, \tau^2 I)$

· posterior $\propto \exp\left[-\frac{1}{2}(y-\theta)^T I(y-\theta)\right] (\tau^2)^{-p/2} \exp\left[-\frac{1}{2\tau^2}\theta^T I\theta\right]$

$= (\tau^2)^{-p/2} \exp\left[-\frac{1}{2}\left\{\theta^T(I + \frac{1}{\tau^2}I)\theta - 2\theta^T y\right\}\right]$

$\sim N(m, V)$ with $V = (1 + 1/\tau^2)^{-1} I = \frac{1}{(1 + 1/\tau^2)} I$, $m = \left(\frac{1}{1 + 1/\tau^2}\right) y$

· "empirical" bc estimate $m = v$ using $\hat{\tau}^2$ from data.

· m (posterior mean) is a shrinkage estimator. If plug in $\hat{\tau}^2_{mle}$, get James Stein.

Bias-Variance Tradeoff Proof:

- Let $\hat{\theta}$ be an estimator of θ.
- $MSE = E[(\hat{\theta}-\theta)^2]$
- $Var(\hat{\theta}) = E[(\hat{\theta}-E(\hat{\theta}))^2]$
- $Bias(\hat{\theta}) = E(\hat{\theta})-\theta$

- Prove that $MSE = Var + Bias^2$

- $MSE = E[(\hat{\theta}-\theta)^2] = E[(\hat{\theta}-E(\hat{\theta}) + E(\hat{\theta})-\theta)^2]$

$$= \underbrace{E[(\hat{\theta}-E(\hat{\theta}))^2]}_{Var} + \underbrace{E[(E(\hat{\theta})-\theta)^2]}_{=(E(\hat{\theta})-\theta)^2} + \underbrace{E[(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)]}_{=0 \text{ bc } E[\hat{\theta}-E(\hat{\theta})] = E(\hat{\theta})-E(\hat{\theta}) = 0}$$

$$= Bias^2$$

So $MSE = Var + Bias^2$. ■

Gauss-Markov Proof (Repeat from Linear)

- Prove that $\hat{\beta}_{LS}$ is BLUE among all linear predictors of θ, with "best" = smallest variance.

- Let $a'y$ be an arbitrary linear predictor.
- We know: $E(y)=X\beta$, $Var(y)=\sigma^2 I$, $\hat{\beta} = p'My$ for $\lambda'\beta = p'X\beta$. $\hat{\beta} \sim N(X\beta, \sigma^2(X'X)^{-1})$

- $Var(a'y) = Var(a'y - \hat{\beta} + \hat{\beta}) = Var(\underbrace{a'y - p'My}_{} + p'My)$

$$= \underbrace{Var(a'y-p'My)}_{\geq 0 \text{ bc is a variance}} + \underbrace{Var(p'My)}_{Var(\hat{\beta}_{LS})} + \underbrace{2Cov(a'y-p'My, p'My)}_{=0}$$

$$Cov(a'y-p'My, p'My) = (a'-p'M)\,Cov(y)\,Mp = \sigma^2(a'-p'M)Mp = \sigma^2(a'M-p'M)p$$

Then $E(a'y) = \cancel{\lambda'\beta = p'X\beta}$ ($p'My$) bc unbiased, so $a' = p'$, and $(a'M-p'M)=0$.
$E(a'y) = a'E(y) = \cancel{a'X\beta}$ ($a'E(y)$) ↓ $a'M = p'M$

Therefore, $Var(a'y) \geq Var(\hat{\beta}_{LS})$ and $\hat{\beta}_{LS}$ is BLUE.

Proof that Training Error Always Less Than Test Error:

Let: · Expected test error = $R_{TE} = E_{y_{tr}y_{te}}[\sum_{i=1}^{n}(y_i^{te} - x_i'\hat{\beta})^2] = E_{te,tr}[(y_{te}-X\hat{\beta})^T(y_{te}-X\hat{\beta})]$

· Expected training error = $R_{TR} = E_{tr}[(y_{tr}-X\hat{\beta})^T(y_{tr}-X\hat{\beta})]$ where $\hat{\beta} = (X'X)^{-1}X'y_{tr}$

· Built model on y_{tr}, then use same $X, \hat{\beta}$ in test model for new y_{te}.
· Goal: Show $R_{TE} > R_{TR}$.

1) Expand R_{TR}: $R_{TR} = E_{tr}(y_{tr}'y_{tr} - 2y_{tr}'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}) = E_{tr}(y_{tr}'y_{tr}) - E_{tr}(2y_{tr}'X\hat{\beta}) + E_{tr}(\hat{\beta}'X'X\hat{\beta})$

2) Expand R_{TE}: $R_{TE} = E_{te,tr}(y_{te}'y_{te} - 2y_{te}'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}) = E_{te}(y_{te}'y_{te}) - 2E_{te}(y_{te}')E_{tr}(X\hat{\beta}) + E_{tr}(\hat{\beta}'X'X\hat{\beta})$

3) $R_{TR} - R_{TE} = E_{tr}(\cancel{y_{tr}'y_{tr}}) - E_{tr}(2y_{tr}'X\hat{\beta}) + E_{tr}(\cancel{\hat{\beta}'X'X\hat{\beta}}) - \{E_{te}(\cancel{y_{te}'y_{te}}) - 2E_{te}(y_{te}')E_{tr}(X\hat{\beta}) + E_{tr}(\cancel{\hat{\beta}'X'X\hat{\beta}})\}$

$$= -2E_{tr}(y_{tr}'X\hat{\beta}) - 2E_{te}(y_{te}')E_{tr}(X\hat{\beta})$$

This has form $Cov(u,v) = E(uv') - E(u)E(v)^T$, which $= E[(u-E(u))(v-E(v))^T]$
· $E(y) = X\beta$
· $E(X\hat{\beta}) = XE(\hat{\beta}) = X(X'X)^{-1}X'E(y) = X(X'X)^{-1}X'X\beta = X\beta$

$R_{TR} - R_{TE} = -2E[(X\hat{\beta}-X\beta)(y-X\beta)^T] = -2E[(\underbrace{X(X'X)^{-1}X'}_{M}y - X\beta)(y-X\beta)^T]$

$= -2E[(My-X\beta)(y-X\beta)^T]$

$= -2tr\{E[(y-X\beta)M(y-X\beta)^T]\} = -2tr\{E[M(y-X\beta)(y-X\beta)^T]\}$

and $E(y-X\beta)(y-X\beta)' = Cov(y-X\beta)$. Since $y \sim N(X\beta, \sigma^2 I)$, $y-X\beta \sim N(0, \sigma^2 I)$

$= -2\sigma^2 tr(M) = -2\sigma^2 p$

Therefore, $R_{TR} - R_{TE} = -2\sigma^2 p$, so $R_{TE} > R_{TR}$. ■

Model Selection:

1. AIC = $-2\log L + 2p$ — Favors models w/ more params.
2. BIC = $-2\log L + p\log n$ — Favors models w/ fewer params (larger penalty than AIC)

§

- Use for comparing models w. same # of params.
- Bad for figuring out how many params to include in model. For this, need Fwd Selection, Backward Selection, CV or LOOCV.

3. Best Subset: Look at all possible models w/ k features. Use AIC/BIC to compare.
   This is not realistic bc $\binom{n}{k}$ models, and still need to know k, # of params.

4. Forward Selection:
   - Begin w intercept-only model.
   - Fit all models w/ 1 param; keep term w/ most significant β.
   - Keep adding terms 1 by 1, until no more sig. predictors to add.
   - Computationally expensive.

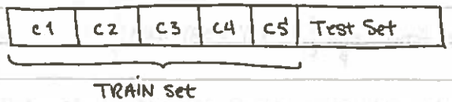
5. Backward Selection:
   - Like forward, but begin with full model, remove least sig. predictors 1 by 1, until no more insignificant predictors to remove.

6. Cross-Validation (CV):
   - Use CV to pick k, the # of params to include.
   - Also good for tuning parameters.

Cross-Validation:

- Use to estimate best # of predictors in a model. Est. test error for ea. k.

1. Split data randomly. 5 or 10 folds common.



2. Pick a model size, say k=6.
   a. For fold C1, fit best model with k=6 params. Pick using best subset. Fit model using C2-C5.
   b. Calc $RSS^{(-C1)} = (Y-X\hat{\beta})^T(Y-X\hat{\beta})$ using C1 as test data.
   c. Repeat a, b for other folds, so have $RSS^{(-C1)}, \ldots, RSS^{(-C5)}$
   d. $RSS_6 = \text{avg}(RSS^{(-C1)}, \ldots, RSS^{(-C5)})$, where RSS_6 is the estimated RSS for k=6.

3. Repeat (2) for all other model sizes you're interested in.

4. Pick the model size w. the smallest RSS. Then estimate test error by fitting this model to the test set, ie calculate RSS_{test}. Note: fit the model with k best params on ALL training data to get $\hat{\beta}$, then $RSS_{test} = (Y_{te}-X\hat{\beta})^T(Y_{te}-X\hat{\beta})$.

Caution! Wrong: 1. Screen for a good set of predictors.
2. Build a model using this subset.
3. Use CV to estimate any unknown tuning params & test error for final model.

Predictors have an unfair advantage! They were chosen in 1, seeing all the data.

- LOOCV: Same as above, but do n folds, leaving out 1 obs ea. time.

- Source of CV Bias:
  - We decreased n by removing a fold, so not fitting model using full n. (Want to estimate the test error using $\hat{\beta}$s fitted w/ all n points.)
  - Can reduce bias with LOOCV, but increases var bc sets (folds) strongly correlated, since they share almost all data points.

Ridge Regression (Repeat from Linear)

- Like OLS, but adds $\ell 2$ norm penalty to diagonal of $X'X$ before inversion.

- loss fcn: $\min_\beta \left\{ (y-X\beta)'(y-X\beta) + \lambda\beta'\beta \right\} = \min_\beta \left\{ \sum_{i=1}^n (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$

- Scale and center data; not invariant to covariates w differing scales.

- Closed-form soln: $\hat\beta_R = (X'X + \lambda I)^{-1} X'y$

- Motivation: Originally proposed to handle collinearity. If X not full rank, or close to be strongly corr predictors, adding λ to diag makes $X'X$ invertible.

- MSE Intuition: $MSE = E[(\hat\beta-\beta)^T(\hat\beta-\beta)] = tr[\sigma^2(X'X)^{-1}]$ since
  - $E(x^TQx) = tr(Q\Sigma) + \mu'Q\mu$
  - $E(\hat\beta-\beta) = 0$
  - $Q = I$
  - $Cov(\hat\beta) = \sigma^2(X'X)^{-1}$

  $= \sigma^2 \sum_{j=1}^p 1/\phi_j^2$ where ϕ_j^2 = eigvals.

  - If any eigvals too close to 0, MSE blows up.
  - Adding λ, a penalty, to RR $\rightarrow$ $MSE = \sigma^2 \sum_{j=1}^p \frac{1}{\phi_j^2+\lambda}$, prevents explosion.

- Bayesian Interpretation:
  - Lhood: $p(y|\beta,\sigma^2) \sim N(X\beta, \sigma^2 I)$ Then $\hat\beta_R$ = posterior mean
    $p(\beta|\sigma^2) \sim N(0, \frac{\sigma^2}{\lambda} I)$

- SVD connection: $X = UDV^T$ as usual. $\hat\beta_R = (X'X+\lambda I)^{-1}X'y$ $\quad X' = VDU^T$

$$\Rightarrow \hat y = X\hat\beta = UDV^T(VDU^TUDV^T + \lambda I)^{-1}VDU^Ty = UD(D^2+\lambda I)^{-1}DU^Ty$$

$$= \frac{UD^2U^Ty}{D^2+\lambda I} = \sum_{j=1}^p u_j \underbrace{\left(\frac{\phi_j}{\phi_j+\lambda}\right)}_{\text{scalar, } \phi_j = \text{eigvals bc } D=\text{diag}(\lambda_j),\ \lambda_j^2=\phi_j} u_j^Ty$$

- So ridge computes coords of y wrt orthonormal basis U (like OLS), but also shrinks by factors $\left(\frac{\phi_j}{\phi_j+\lambda}\right)$, so $\uparrow$ shrinkage applied to basis vecs w/ smaller eigvals.

- Tuning penalty λ: Use CV to tune λ.

Generalized Linear Models:

1. Logistic regression $\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = X\beta + e$ - ie $\phi(\mu) = X\beta+e$; $\phi(y) = \text{logit}(y)$
   - models log-odds
   - can fit using IRLS (equiv to Newton's Method), sgd, etc.
   - $p(y=1|x_i,\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$, call it "$p$". Then $\nabla \log\ell(\beta) = \sum_{i=1}^n x_i(y_i-p)$.
   - for y = binary response only
   - Can set up as a latent variable model, with latent $z_i = X\beta+e$. Similar to probit approach.

   * "link" fctn, logit in this case, describes relationship btwn expected value of the response and the linear prediction of the covariates

2. Poisson glm $E(y|x) = e^{\theta'x}$, so model is $\log\theta = X\beta+e$, where y = counts
   - e^{β_j} represents effect of a 1-unit increase in x on θ.

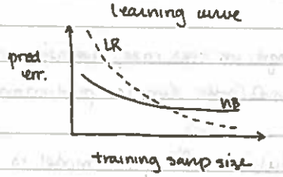
Naive Bayes

- Classify $y \in \{0,1\}$ binary points using Bayes Rule.
- "Naive" bc uses assumption of conditional independence.

- Bayes Rule: $P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)}$

- Reduces # of predictors by assuming condit indep
  - Regular Bayes Rule needs $2(2^k - 1) + 1$ where k = # predictors.
    (# y levels; +1 for $p(y)$ prior)
  - Naive Bayes only needs $(2k+1)$ params estimated.

- NB vs Logistic Regression:
  - LR is a ↑ complex model - takes longer to converge.



- Example on Discrete X: (HW 3, 2.2)
- prior: $P(Y) = .5$ Say have 2 binary covariates, $X_1 - X_2$
- given: $P(X_1=1|Y=1) = .8$ $P(X_1=0|Y=0) = .7$
  $P(X_2=1|Y=1) = .5$ $P(X_2=0|Y=0) = .9$

(1) Calc Decision rule for Y, for every combo of (X_1, X_2). If $P(Y=1|X_1,X_2) > .5$, classify as $Y=1$.

Ex: $P(Y=1|X_1=0,X_2=0) = \frac{P(X_1=0|Y=1)\,P(X_2=0|Y=1)\,P(Y=1)}{P(X_1=0|Y=1)P(X_2=0|Y=1)P(Y=1) + P(X_2=0|Y=0)P(X_2=0|Y=0)P(Y=0)}$

$= .2(.5)(.5) / \left[.2(.5)(.5) + .7(.9)(.5)\right] = .14$, so classify as $Y=0$.

Repeat for 3 other (X_1, X_2) combos.

(2) Error rate = sum $(a+b+c+d)$ where $a = P(X_1=0, X_2=0|Y=1)$, ie of seeing these x's w/ opp y.

- $a = P(X_1=0|Y=1)\,P(X_2=0|Y=1)\,P(Y=1) = .5(.2)(.5) = .05$
- For b, c, d, repeat for each combo of x's.

- Example from Midterm:
- $f(x,y)$ = bivariate uniform, ie $f(x,y) = \frac{\mathbb{1}(x_\ell < x < x_h,\ y_\ell < y < y_h)}{(x_h - x_\ell)(y_h - y_\ell)}$

- Data:

x	0	1	1	2	8	6	7	5	6	0
y	2	0	5	3	4	6	4	7	3	6
class	1	1	1	1	1	2	2	2	2	2

a) What are MLEs for x_ℓ, x_h, y_ℓ, y_h? $x_{(1)}, x_{(n)}, y_{(1)}, y_{(n)}$

b) Use Bayes Rule to estimate the point (0,1)

$$\frac{P(\text{class}=1|x=0,y=1)}{P(\text{class}=0|x=0,y=1)} = \frac{\textcircled{1}\ P(X=0,y=1|\text{class}=1)\ P(\text{class}=1)}{\textcircled{2}\ P(X=0,y=1|\text{class}=0)\ P(\text{class}=0)}$$

(.9 over $P(\text{class}=1)$; .5 under $P(\text{class}=0)$)

① $f(x,y|\text{class}=1) = \frac{\mathbb{1}(0<x<8, 0<y<5)}{(8-0)(5-0)} = \frac{1}{40} = 1/40$

② $f(x,y|\text{class}=0) = \frac{\mathbb{1}(0<x<7, 3<x<7)}{(7-0)(7-3)} = \frac{1}{28} = 0$

- (0,1) outside support of class 2, so classified as class 1.

(c) Use Bayes Rule to estimate point (2,4)

$\frac{P(\text{class}=1|x=2,y=4)}{P(\text{class}=2|x=2,y=4)} = \frac{1/40}{1/28} = \frac{28}{40} = 7/10 < 1$, so classify as class 2.

K Means Clustering:

- Given number of clusters k, want to classify obs $y = y_1 \dots y_n$ into clusters.
- No y labels; unsupervised.

- Objective fcn: $\min_{\{\mu_1 \dots \mu_k\}} \sum_{i=1}^{n} \| x - \mu_i \|^2$, ie want to divide obs into sets which minimize the within-cluster variance.
  - μ_k = cluster mean k.

- Algorithm:
  - Randomly set k cluster centers to begin.
  - Assign every point to its closest cluster.
  - Recalc cluster mean as avg of all points in clusters.
  - Repeat.

- Ex: Say have n=7 points, $\{-41, -2, -1, 0, 3, 4, 5\}$. Use K=2, with initial centers at $\mu_1 = -1$, $\mu_2 = 4$.

1) -41 -2 -1 0 3 4 5
   (-41, -2, -1, 0: closer to -1; 3, 4, 5: closer to 4)

New centers: -11, 4

2) -41 -2 -1 0 3 4 5
   (-41: closest to -11; -2, -1, 0, 3, 4, 5: closest to 4)

New centers: -41, $3/2$.

No changes in future iters, so we are done.

Robust Statistics:

1. Estimating location parameters w outliers:
   - α-trimmed mean: Trim α proportion of min/max obs from ea tail, then compute the mean.
   - α-Winsorized mean: Like α-trimmed, but replaces the points in the tails with the most extreme non-tail points, then calc mean.

   - Ex: Data set is $\{2, 4, 5, 10, 100\}$. 20% trimmed mean: $(4+5+10)/3 = 6.33$
     20% windsor mean: $(4+4+5+10+10)/5 = 6.6$

2. Sensitivity Curve: Measures effect of changing 1 data point in the set.
   - T = the sample statistic
   - $SC_n(x, T) = \dfrac{T(x_1, \dots, x_{n-1}, x_{new}) - T(x_1 \dots x_{n-1})}{1/n}$

3. Influence Fcn: $\lim_{n \to \infty} SC_n(x, T)$
   - If bounded (not $-\infty$ or ∞), we call the statistic T 'robust'.

Ex: With prob $(1-\alpha)$, points from $U(0, \theta)$. With prob α, points from other dist, ie noise.

(a) Say $\alpha = 0$. What is mle of θ? $\hat{\theta} = x_{(n)}$, the sample max.

(b) Sensitivity curve for $\hat{\theta}$: $SC_n(\hat{\theta}, x) = n[\max(x_1 \dots x_{n-1}, x_{new}) - \max(x_1 \dots x_{n-1})]$

$= n[\max\{x_{(n)}, x_{new}\} - x_{(n)}]$

(c) Based on SC_n, is $\hat{\theta}$ appropriate if there are outliers, ie $\alpha > 0$?

No. If set $x_{new} = \infty$, $SC_n = n[\infty - x_n] = \infty$

(d) If know $\alpha = .1$, suggest a robust variant of $\hat{\theta}$.

α-Trimmed sample max. Trim α off of top of sorted data set (and bottom, though won't make a difference here). Then compute sample max.

M-Estimators:

- Generalizes MLEs. Instead of setting $\frac{\partial}{\partial\theta}(\log L)=0$ & solving for $\hat\theta$, we instead minimize some loss fctn $\rho(x_i,\theta)$.

- Instead of "score fctn" $\frac{\partial}{\partial\theta}\log L=0$, use $\psi(z)=\frac{\partial\,\rho(z)}{\partial z}$

↓

- Setting is $\hat\theta=\underset{\theta}{\text{argmin}}\left\{\sum_{i=1}^n \rho(x_i,\theta)\right\}$ → solve for $\hat\theta$ using $\sum_{i=1}^n \psi(x_i,\theta)=0$

Example: when θ = mean μ, use squared loss fctn $(x_i-\theta)^2$.

- $\rho(x_i,\theta)=(x_i-\theta)^2$, so $\psi(x,\theta)=-2(x_i-\theta)$ → $\sum_{i=1}^n -2(x_i-\theta)=0$ → $\Sigma x_i - n\theta=0$ → $\hat\theta=\bar x$

Example: when θ = median, use absolute loss fctn $|x_i-\theta|$.

- $\rho(x_i,\theta)=|x-\theta|$, so $\psi(x,\theta)=\text{sign}(x-\theta)$. Then $\sum_{i=1}^n \text{sign}(x-\theta)=0$ → $\hat\theta$ = sample median $\tilde x$

⊛ if $\psi(x,\theta)$ bounded for all x, then $\hat\theta$ is a robust estimator.

Asymptotic Relative Efficiency:

- Ratio of asymptotic means of two estimators.
- Estimator w. smaller variance is more efficient.

Ex: Mean vs. Median. $\bar x \sim N(\mu, \sigma^2/n)$

$\tilde x \sim N(M_n, \frac{\sigma^2\pi}{2n})$

- $ARE(\bar x,\tilde x)=\frac{\sigma^2\pi/2n}{\sigma^2/n}=\frac{\pi/2}{1}=\frac{\pi}{2}>1$, so $\bar x$ more efficient than $\tilde x$.

EM Algorithm: For A 2-Component Gaussian Mixture

- $f(x|\theta)=\pi N(x|\mu_1,\sigma^2)+(1-\pi)N(x|\mu_2,\sigma^2)$ (Say σ^2 known)

  $\theta=\{\pi,\mu_1,\mu_2\}$

- Introduce latent z_i vars to indicate class membership for each obs x_i. $z_i=1$ or 0.

- Original lhood: $\prod_{i=1}^n\left[\pi\underbrace{N(x_i|\mu_1,\sigma^2)}_{\phi_1(x_i)}+(1-\pi)\underbrace{N(x_i|\mu_2,\sigma^2)}_{\phi_2(x_i)}\right]=\prod_{i=1}^n\left[\pi\phi_1(x_i)+(1-\pi)\phi_2(x_i)\right]$

The latent z_i's allow us to eliminate the summation in the lhood.

- Augmented lhood: $\prod_{i=1}^n \underbrace{\phi_1(x_i)^{z_i}\phi_2(x_i)^{(1-z_i)}}_{f(x_i|z_i,\theta)}\underbrace{\pi^{z_i}(1-\pi)^{(1-z_i)}}_{p(z_i)\sim\text{Bernoulli}(\pi)}$ so $\prod_{i=1}^n f(x_i|z_i,\theta)\,p(z_i)$

- LogL Augmented: $\sum_{i=1}^n\left[z_i\log\phi_1(x_i)+(1-z_i)\log\phi_2(x_i)+z_i\log\pi+(1-z_i)\log(1-\pi)\right]$ (logLA)

E Step Using current params, update z_i's, using their expected values.

$$P(z_i=1)=\gamma_i=E(z_i|x_i,\theta)=\frac{\pi\,\phi_1(x_i)}{\pi\,\phi_1(x_i)+(1-\pi)\phi_2(x_i)}$$

Then classify $z_i=1$ if $P(z_i=1)>.5$, else $z_i=0$.

M Step Using updated z_i's, estimate updated MLEs for each param (π,μ_1,μ_2).

$(\hat\pi)$ $\frac{\partial}{\partial\pi}\log LA=0$ → $\frac{\sum z_i}{\pi}-\frac{\sum(1-z_i)}{1-\pi}=0$ → $\frac{\pi}{1-\pi}=\frac{\sum z_i}{n-\sum z_i}\overset{\text{div by } n}{\to}\frac{\pi}{1-\pi}=\frac{\sum z_i/n}{1-\sum z_i/n}$

So $\hat\pi=\frac{\sum z_i}{n}$ where $z_i=1$ or 0 based on E step updates.

$(\hat\mu_1)$ $\frac{\partial}{\partial\mu_1}\log LA=0$ → $\sum_{i=1}^n -2z_i(x_i-\mu_1)=0$ → $2\Sigma z_i x_i-2\Sigma z_i\mu_1=0$ → $\hat\mu_1=\frac{\sum x_i z_i}{\sum z_i}$

which is the mean of the obs in updated group 1.

- Comes from lhood $A=\prod_{i=1}^n(\sigma^2)^{-z_i/2}\exp\left[\frac{-1}{2\sigma^2}(x_i-\mu_1)^2\right]^{z_i}(\sigma^2)^{-(1-z_i)/2}\exp\left[\frac{-1}{2\sigma^2}(x_i-\mu_2)\right]^{(1-z_i)}\pi^{z_i}(1-\pi)^{(1-z_i)}$

$(\hat\mu_2)$ $\frac{\partial}{\partial\mu_2}\log LA=0$ → $\sum_{i=1}^n -2(1-z_i)(x_i-\mu_2)=0$ → $\sum_{i=1}^n(1-z_i)x_i-\sum_{i=1}^n(1-z_i)\mu_2=0$ → $\hat\mu_2=\frac{\sum_{i=1}^n(1-z_i)x_i}{\sum_{i=1}^n(1-z_i)}$

EM-Related Exponential Example

- Say there are 2 sets of (independent) bulbs.
- $\{y_1 \dots y_n\}$ We observe these the whole time, so know failure_times.
- $\{E_1 \dots E_m\}$ We do not observe failure times, latent $\{z_1 \dots z_m\}$ - just whether bulbs in set 2 are on or off at time T. $\{E_1 \dots E_m\}$ are 0 or 1, 1 = on at T.

- $y \sim Exp(\theta)$, $z \sim Exp(\theta)$, where $f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}$. So $F(x|\theta) = 1 - e^{-x/\theta}$

- Goal: estimate $\hat{\theta}$. If we had z's, $\hat{\theta} = \frac{\left(\sum_{i=1}^{n} y_i + \sum_{j=1}^{m} z_i\right)}{n+m}$

- Since don't have z's, we can replace them with their expected values, $E(z_i | E_i, \theta)$.
  - Each z_i only depends on its own E_i and θ.
  - E_i is 0 or 1 only:
    - If $E_i = 1$, $E(z_i | E_i, \theta) = T + \theta$, since mean of $Exp(\theta) = \theta$, and exponential dist. is memoryless.
    - If $E_i = 0$, we can solve for this:

$$E(z_i) = \underbrace{E[z_i | E_i = 0, \theta]}_{?} \underbrace{P(E_i = 0)}_{\text{prob of failure before T, ie cdf eval at T: } 1 - e^{-T/\theta}} + \underbrace{E[z_i | E_i = 1, \theta]}_{T+\theta \text{ from above}} \underbrace{P(E_i = 1)}_{\text{prob of failure after time T, ie 1-cdf eval at T: } e^{-T/\theta}}$$

($E(z_i)$: θ bc mean of $Exp(\theta)$)

$$\theta = E[z_i | E_i = 0, \theta]\,(1 - e^{-T/\theta}) + (T+\theta)\, e^{-T/\theta}$$

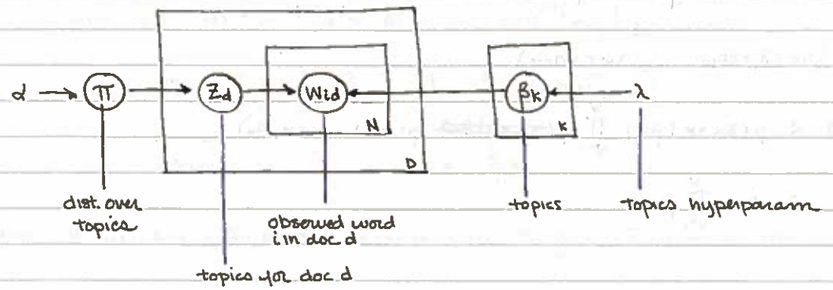
$$E[z_i | E_i = 0, \theta] = \frac{\theta - (T+\theta) e^{-T/\theta}}{1 - e^{-T/\theta}}$$

- Plug the expressions for $E[z_i | E_i = 1, \theta]$ and $E[z_i | E_i = 0, \theta]$ into the expression for $\hat{\theta}$, in place of z_i's, based on δ_i for each obs.

Topic Modeling Review

- Goal: Classify a document into a topic based on its words. Assume naive: 1 topic per doc

- How model is generated:
  - Randomly pick a topic z_d, from dist. over topics $\pi(z_1 \dots z_K) \sim Dir(\alpha_1 \dots \alpha_K)$
  - Topic z_d is just a distribution over words, fully specified by probs of words, $\beta = (\beta_1 \dots \beta_V)$ where V = # of words in vocab.
  - Words for doc d are randomly selected from vocab with probs β.



- z_d = topic of document d, for d in $\{1 \dots N\}$ (N total docs to classify)
- $z_d \in \{1 \dots K\}$, where K = # of topics

- $\pi \sim Dir(\alpha_1 \dots \alpha_K)$ is dist. of topics, so $P(z_d = k) = \pi_k$ for d in $1 \dots N$

- w_{id} = word i in doc d. $P(w_{id} | z_d = k) = \beta_{kw}$

- $\beta_k = (\beta_{k1} \dots \beta_{kV}) \sim Dir(\lambda_1 \dots \lambda_{nd})$ is probs for each word in vocab, in topic k. (Each topic has its own dist. of words.)

- Can derive full conditionals + use a Gibbs Sampler.

① $p(\pi) \propto p(\pi) \cdot \prod_{d=1}^{D} p(z_d = k | \pi)$ with $p(\pi) \sim Dir(\alpha_1 \dots \alpha_K)$ and $p(z_d = k) = \pi_k$

$$= \pi_1^{\alpha_1 - 1} \dots \pi_K^{\alpha_K - 1} \cdot \prod_{d=1}^{D} \pi_1^{\mathbb{1}(z_d = 1)} \dots \pi_K^{\mathbb{1}(z_d = K)}$$

$$= \pi_1^{\alpha_1 - 1} \dots \pi_K^{\alpha_K - 1} \cdot \pi_1^{\sum_{d=1}^{D} \mathbb{1}(z_d = 1)} \dots \pi_K^{\sum_{d=1}^{D} \mathbb{1}(z_d = K)}$$

Let $\sum_{d=1}^{D} \mathbb{1}(z_d = k) = n_k$, # docs classed as topic k.

$$= \pi_1^{n_1 + \alpha_1 - 1} \dots \pi_K^{n_K + \alpha_K - 1} \sim Dir(n_1 + \alpha_1, \dots, n_K + \alpha_K)$$

② $p(\beta_k|\dots) \propto p(\beta_k) \prod_{d=1}^{D} \prod_{i=1}^{n_d} p(w_{id}|\beta_k)$

$= \beta_{k1}^{\lambda_1-1} \cdots \beta_{kV}^{\lambda_V-1} \cdot \prod_{d=1}^{D} \prod_{i=1}^{n_i} \prod_{v=1}^{V} \beta_{kv}^{\mathbb{1}(w_{id}=v)}$

$= \beta_{k1}^{\lambda_1-1} \cdots \beta_{kV}^{\lambda_V-1} \cdot \prod_{v=1}^{V} \beta_{kv}^{\sum_d \sum_i \mathbb{1}(w_{id}=v)}$ $= m_{kv}$, the # of times word v appears in topic k, across all docs.

$= \beta_{k1}^{\lambda_1-1} \cdots \beta_{kV}^{\lambda_V-1} \cdot \beta_{k1}^{m_{k1}} \cdots \beta_{kV}^{m_{kV}}$

$\sim Dir(\lambda_1 + m_{k1}, \dots, \lambda_V + m_{kV})$

③ $p(z_d=k|\dots) \propto p(z_d=k|\beta_k) \prod_{v=1}^{V} p(w_{id}|z_d=k, \beta_k)$

$= \pi_k \cdot \prod_{v=1}^{V} \beta_{kv}$

Sufficient Statistics

1. Factorization Theorem: Let $f(x|\theta)$ be the joint lhood of sample $x_1 \dots x_n$. $T(x)$ is a sufficient statistic for θ iff can factor $f(x|\theta)$ as

$$f(x|\theta) = g(T(x)|\theta) \cdot h(x)$$

Useful for finding a sufficient stat.

2. Ex 1, Normal Mean: $x_1 \dots x_n \overset{iid}{\sim} N(\theta, \sigma^2)$. Find $T(x)$, a sufficient stat for θ.

$$f(x|\theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-1}{2\sigma^2}(x_i-\theta)^2\right] = (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$$

$$= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right]}_{h(x)} \cdot \exp\left[\frac{-1}{2\sigma^2}\left\{n\theta^2 - 2\theta\sum_{i=1}^{n} x_i\right\}\right]$$

$T(x) = \sum_{i=1}^{n} x_i$

- Note: If $T(x)$ is a sufficient stat, $c(x)$ is also. So $\frac{1}{n}T(x) = \bar{x}$ is a sufficient stat.

3. Bernoulli Ex: $x_1 \dots x_n \sim$ Bernoulli(p). Find a sufficient stat for p.

$$f(x|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i} \quad \text{so } h(x)=1, \quad T(x) = \sum_{i=1}^{n} x_i$$

4. Weibull Ex: β is a parameter, γ is known.

$$f(x|\beta) = \prod_{i=1}^{n} \frac{\gamma}{\beta} x^{\gamma-1} \exp[-x^\gamma/\beta] = \left(\frac{\gamma}{\beta}\right)^n \exp\left[-\sum x_i^\gamma/\beta\right] \cdot \prod_{i=1}^{n} x_i^{\gamma-1}$$

$\cdot h(x) = \prod_{i=1}^{n} x_i^{\gamma-1}$ and $T(x) = \sum_{i=1}^{n} x_i^\gamma$

⊛ For exponential families, $f(x|\underline{\theta}) = h(x)\, c(\underline{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\underline{\theta})\, t_i(x)\right)$

Then $T(x) = \left[\sum_{i=1}^{n} t_1(x_i), \dots, \sum_{i=1}^{n} t_k(x_i)\right]$ is a suff stat for $\underline{\theta}$.

Minimal Sufficient Statistic:

- A sufficient statistic $T(x)$ is minimal if for any other stat $\tilde{T}(x)$, $T(x)$ is a fctn of $\tilde{T}(x)$.
- Ex: $\bar{x}$ is the min suff stat. The whole data set, $x_1 \dots x_n$, is a suff stat.
  $\bar{x}$ is a fctn of $x_1 \dots x_n$

- $c\,T(x)$ is a min suff stat if $T(x)$ is, where c = constant

- How to find a min. suff. stat? Thm 6.2.13 (Berger + Casella).
  Say $\exists\ T(x)$ so that $f(x|\theta) / f(y|\theta)$ is constant wrt θ. Then $T(x)$ is min suff stat.

Ex: Find a min suff stat for Weibull, for β, where γ is known.

$$f(x|\beta) = \prod_{i=1}^{n} \frac{\gamma}{\beta} x^{\gamma-1} \exp[-x^\gamma/\beta] = \left(\frac{\gamma}{\beta}\right)^n \exp\left[-\sum_{i=1}^{n} x_i^\gamma/\beta\right] \prod_{i=1}^{n} x_i^{\gamma-1}$$

Ratio: $\dfrac{\left(\gamma/\beta\right)^n \exp\left[-\sum_{i=1}^{n} x_i^\gamma/\beta\right] \prod_{i=1}^{n} x_i^{\gamma-1}}{\left(\gamma/\beta\right)^n \exp\left[-\sum_{i=1}^{n} y_i^\gamma/\beta\right] \prod_{i=1}^{n} y_i^{\gamma-1}}$

- Need $\sum x_i^\gamma = \sum y_i^\gamma$ so β's cancel, ie ratio constant wrt β.
- $T(x) = \sum x_i^\gamma$ is min suff stat.

Ex: Find a min sufficient stat for θ in $U[0,\theta]$.

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{\prod_{i=1}^{n} \frac{1}{1-\theta} I_{[0,\theta]}(x_i)}{\prod_{i=1}^{n} \frac{1}{1-\theta} I_{[0,\theta]}(y_i)} = \frac{\left(1/\theta^n\right) I_{(x)[0,\theta]}}{\left(1/\theta^n\right) I_{(y)[0,\theta]}}$$

- Can split indicators:
- If $\max(x) > \theta$, then all $x < \theta$. If $\min(x) > 0$, all $x > 0$, but doesn't depend on θ.

$$\frac{\prod_{i=1}^{n} I_{(-\infty,\theta]}(x_i)\ I_{[0,\infty)}(x_i)}{\prod_{i=1}^{n} I_{(-\infty,\theta]}(y_i)\ I_{[0,\infty)}(y_i)} = \frac{I_{(-\infty,\theta]}(x_{(n)})}{I_{(-\infty,\theta]}(y_{(n)})}$$ so $x_{(n)}$ is min suff stat.

Chapter 1 - Basics

- Calculus Review: U-substitution

1) $\int (2x+3)(x^2+3x)^7 dx$ Let $u = x^2+3x$, so $du = (2x+3)dx$. $= \int u^7 du = \frac{u^8}{8} = \frac{(x^2+3x)^8}{8}$

2) $\int (3-x)^{10} dx$ Let $u = 3-x$, $du = -dx$. $= -\int u^{10} du = \frac{-u^{11}}{11} = \frac{-(3-x)^{11}}{11}$

3) $\int (7x+9)^{1/2} dx$ Let $u = 7x+9$, $du = 7dx$. $= \frac{1}{7}\int u^{1/2} du = \frac{1}{7} \cdot \frac{u^{3/2}}{3/2} = \frac{2}{21} u^{3/2} = \frac{2(7x+9)^{3/2}}{21}$

4) $\int x^3 (1+x^4)^{-1/3} dx$ Let $u = 1+x^4$, $du = 4x^3 dx$. $= \int \frac{1}{4} u^{-1/3} du = \frac{1}{4} \frac{u^{2/3}}{(2/3)} = \frac{3}{8} u^{2/3} = \frac{3(1+x^4)^{2/3}}{8}$

5) $\int e^{5x+2} dx$. Let $u = 5x+2$, $du = 5dx$. $= \frac{1}{5}\int e^u du = \frac{1}{5} e^u = \frac{1}{5} e^{5x+2}$

6) $\int 4\cos(3x) dx$. Let $u = 3x$, $du = 3dx$. $= \int \frac{4}{3}\cos(u) du = \frac{4}{3}\sin(u) = \frac{4}{3}\sin(3x)$

7) $\int \frac{\sin(\ln(x))}{x} dx$. Let $u = \ln(x)$, $du = \frac{1}{x}dx$. $= \int \sin(u) du = -\cos(u) = -\cos(\ln(x))$.

- Calculus Review: Integration by Parts $\int u\,dv = uv - \int v\,du$

1) $\int xe^x dx$. Let $u = x$, $dv = e^x dx$. Then $du = dx$, $v = \int e^x dx = e^x$.

$= xe^x - \int e^x dx = xe^x - e^x$

2) $\int x\ln(x) dx$. Let $u = \ln(x)$, $dv = x\,dx$. Then $du = \frac{1}{x}dx$, $v = \int x\,dx = x^2/2$

$= uv - \int v\,du = \frac{x^2}{2}\ln(x) - \int \frac{x^2}{2}\left(\frac{1}{x}\right) dx = \frac{x^2}{2}\ln(x) - \int \frac{x}{2} dx = \frac{x^2}{2}\ln(x) - \frac{x^2}{4}$

3) $\int x\cos(3x) dx$. Let $u = x$, $dv = \cos(3x)dx$. Then $du = dx$, $v = \int \cos(3x) dx$. Let $u' = 3x$, $du' = 3dx$,

so $v = \int \frac{1}{3}\cos(u') du' = \frac{1}{3}\sin(u') = \frac{1}{3}\sin(3x)$.

Then $uv - \int v\,du = \frac{x}{3}\sin(3x) - \underbrace{\int \frac{1}{3}\sin(3x) dx}_{\text{another u-sub}} = \frac{x}{3} + \left(\frac{1}{9}\right)\cos(3x)$

4) $\int \frac{\ln(x)}{x^5}dx = \int \ln(x)\cdot x^{-5}dx$. Let $u=\ln(x)$, $dv=x^{-5}dx$. Then $du=\frac{1}{x}dx$, $v=\int x^{-5}dx=\frac{-x^{-4}}{4}$

$= uv-\int vdu = -\frac{\ln(x)}{4x^4} - \int -\frac{x^{-4}}{4}\left(\frac{1}{x}\right)dx = \frac{-\ln(x)}{4x^4} + \int \frac{1}{4}x^{-5}dx = \frac{-\ln(x)}{4x^4} - \frac{1}{4}\frac{x^{-4}}{4} = \frac{-\ln(x)}{4x^4} - \frac{1}{16x^4}$

5) $\int \ln(x)dx$. Let $u=\ln(x)$, $dv=(1)dx$. Then $du=\frac{1}{x}dx$, $v=\int 1dx = x$.

$uv-\int vdu = x\ln(x) - \int x\left(\frac{1}{x}\right)dx = x\ln(x) - \int(1)dx = x\ln(x) - x$.

6) $\int x^2e^{3x}dx$. Let $u=x^2$, $dv=e^{3x}dx$. Then $du=2xdx$, $v=\int e^{3x}dx = \frac{e^{3x}}{3}$

$uv-\int vdu = \frac{x^2e^{3x}}{3} - \underbrace{\int \frac{e^{3x}}{3}(2x)dx}$

Needs a second int. by parts:

Let $u=2x$, $dv=\frac{e^{3x}}{3}dx$, then $du=2dx$, $v=\int\frac{e^{3x}}{3}dx = \frac{e^{3x}}{9}$

$uv-\int vdu = \frac{2xe^{3x}}{9} - \int\frac{2}{9}e^{3x}dx = \frac{2xe^{3x}}{9} - \frac{2e^{3x}}{3\cdot 9} = \frac{2xe^{3x}}{9} - \frac{2e^{3x}}{27}$

Put back together: $= \frac{x^2e^{3x}}{3} - \frac{2xe^{3x}}{9} + \frac{2e^{3x}}{27}$

CH 1 Basic Formulas:

1) Bayes Thm: $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$, where $P(X) = \int P(X|Y)P(Y)dy$ or $\sum_i P(X|Y_i)P(Y_i)$ (discrete)

2) Def. of conditional probability: $P(Y|X) = P(X,Y)/P(X)$

3) To prove a fcn is a pdf:
   1. $f(x)$ is non-negative; $f(x)\geq 0\ \forall x$
   2. $f(x)$ integrates (or sums, for discrete) to 1: $\int_{-\infty}^{\infty} f(x)=1$ or $\sum_x f(x)=1$.

CH2- Transformations + Expectations

- **Thm 2.1.3 - cdf of a transformation:** X has cdf $F_X(x)$. Let $Y=g(X)$.
  - Check if $g(x)$ increasing ($\frac{\partial}{\partial x}g(x)>0$) or decreasing.
  - If $g(x)$ increasing, $F(y)=F_X(g^{-1}(y))$
  - If $g(x)$ decreasing, $F(y)=1-F_X(g^{-1}(y))$

- Ex: $X\sim U(0,1)$. Let $Y=g(X)=-\log(x)$. $y=-\log(x) \rightarrow x=e^{-y} \rightarrow g^{-1}(y)=e^{-y}$
  - Check $g(x)$: $\frac{d}{dx}-\log(x) = \frac{-1}{x} < 0$ since $x\in(0,1)$. So decreasing.
  - $F(y)=1-e^{-y}$, which we recognize as the cdf of an exponential(1).

- **Thm 2.1.5 - Pdf of a transformation:** Let X, and $Y=g(X)$.
  - $f_Y(y)=f_X(g^{-1}(y))|J|$ where $|J|=\left|\frac{d}{dy}g^{-1}(y)\right|$

- Ex: $f(x)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$, $0<x<\infty$. Find pdf of $Y=1/X$. ($X\sim Ga(a,b)$).
  - Find $g^{-1}(y)$: $y=1/x \rightarrow x=1/y$, so $g^{-1}(y)=1/y = y^{-1}$
  - $|J| = |-y^{-2}| = 1/y^2$
  - $f(y)=f_X(g^{-1}(y))|J| = \frac{b^a}{\Gamma(a)}\left(\frac{1}{y}\right)^{a-1}e^{-b(1/y)}\left(\frac{1}{y^2}\right) = \frac{b^a}{\Gamma(a)}y^{-a-1}e^{-by} \sim IG(a,b)$

- Ex 2.1.7, Square Transformation: Let X be continuous, and $Y=X^2$. Find cdf of Y.
  $P(Y\leq y) = P(X^2\leq y) = P(-\sqrt{y}\leq X\leq\sqrt{y}) = P(X\leq\sqrt{y}) - P(X\leq -\sqrt{y})$

- Ex 2.3.9 - Normal-Chisq Relationship: Let $X\sim N(0,1)$. $Y=X^2$. Find pdf of Y.
  - Since $Y=X^2 \rightarrow x=\sqrt{y}, -x=\sqrt{y}$, split into 2 partitions:
  - A_1: $x\in(-\infty,0)$: $y=x^2 \rightarrow x=\sqrt{y}$ so $g^{-1}(y)=\sqrt{y}$.
  - A_2: $x\in(0,\infty)$: $y=x^2 \rightarrow x=-\sqrt{y}$ so $g^{-1}(y)=-\sqrt{y}$
  } Support of y is $y\geq 0$ for both
  - Jacobian: A_1: $|J| = \left|\frac{\partial}{\partial y}g^{-1}(y)\right| = \left|\frac{1}{2}y^{-1/2}\right| = 1/2\sqrt{y}$
    A_2: $|J| = \left|\frac{\partial}{\partial y}g^{-1}(y)\right| = \left|-\frac{1}{2}y^{-1/2}\right| = 1/2\sqrt{y}$
  - pdf: $f(y) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\sqrt{y})^2\right](1/2\sqrt{y}) + \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(-\sqrt{y})^2\right](1/2\sqrt{y})$

$= \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}} + \frac{1}{\sqrt{2\pi}}e^{-y/2}\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{y}}e^{-y/2} \sim \chi^2_1$

Thm 2.10 - Probability Integral Transform Theorem:

- X has cdf $F(x)$. Define Y as $Y = F(x)$. Then $Y \sim U(0,1)$
- Used for random number generation.

2.2: Expected Values:

- $E(g(X)) = \int g(x) f(x)\,dx$, or $\sum_i g(x) f(x)$ for discrete.

- **Ex: Exponential mean.** $X \sim \text{Exp}(\lambda)$. Find $E(X)$. $\quad f(x) = \frac{1}{\lambda} e^{-x/\lambda}$.
  - $E(X) = \int_0^\infty x \left(\frac{1}{\lambda}\right) e^{-x/\lambda}\,dx$. Int by parts: $\int u\,dv = uv - \int v\,du$.

Let $u = x$, $dv = \frac{e^{-x/\lambda}}{1/\lambda}\,dx \longrightarrow du = dx$, $v = -e^{-x/\lambda}$ bc $v = \int \frac{1}{\lambda} e^{-x/\lambda}\,dx = -e^{-x/\lambda}$

Then $E(X) = -x e^{-x/\lambda}\Big|_0^\infty + \underbrace{\int_0^\infty e^{-x/\lambda}\,dx}_{\exp(\lambda) \text{ kernel, int. to } 1/\lambda}$

$\lim_{x\to\infty}(-x e^{-x/\lambda}) - \lim_{x\to 0}(-x e^{-x/\lambda}) + \frac{1}{\lambda} = \frac{1}{\lambda}$, so $E(X) = 1/\lambda$.

- **Ex: Poisson mean:** $Y \sim \text{Pois}(\lambda)$, so $p(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Find $E(Y)$.

- $E(Y) = \sum_{y=0}^\infty \frac{y \lambda^y e^{-\lambda}}{y!}$ Then $y=0$ term is 0, so $= \sum_{y=1}^\infty y \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^\infty \frac{\cancel{y}\,\lambda^y e^{-\lambda}}{\cancel{y}(y-1)!}$

$= \sum_{y=1}^\infty \frac{\lambda^y e^{-\lambda}}{(y-1)!} \longrightarrow \lambda \sum_{y=1}^\infty \frac{\lambda^{y-1} e^{-\lambda}}{(y-1)!}$ Then let $x = y-1$: $\lambda \sum_{x=0}^\infty \frac{\lambda^x e^{-\lambda}}{x!} = \lambda$.

pdf for $X \sim \text{pois}(\lambda)$, sums to 1

- So $E(Y) = \lambda$.

- **Mgfs:**

1) $M_X(t) = E(e^{Xt})$.

2) If $X \perp Y$, $M_{X+Y}(t) = M_X(t) M_Y(t)$. (& if $Z = X+Y$, $M_Z(t) = M_X(t) M_Y(t)$.

3) How to find moments: n^{th} moment of X: $M_X(t)^{(n)} = \frac{d^{(n)}}{dt^{(n)}} M_X(t)$ eval at $t=0$.

(n^{th} deriv of $M_X(t)$ wrt t, eval at $t=0$.

4) $M_{aX+b}(t) = E(e^{(aX+b)t}) = e^{bt} E(e^{Xat}) = e^{bt} M_X(at)$

CHAPTER 3 - Approximations + Relationships

- **Binom Approximation of Poisson:**
  - If $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Pois}(\lambda)$, then $P(X=x) \approx P(Y=x)$ for $\log n$, small np.

- **Normal Approximation to Binomial:**
  - $\text{Binom}(n,p) \approx N(np,\ n\cdot(1-p))$
    - ok if $\min(np,\ n(1-p)) \geq 5$

- Continuity correction: $P(X \leq x) \approx P(Y \leq x + 1/2)$

  $P(X = x) \approx P(Y \geq x - 1/2)$

- **Biased Estimation:**
  - Bias = $E(\hat\theta) - \theta$ for estimator $\hat\theta$ of θ.
  - Can derive unbiased estimators. Example:

- **Ex:** $Y_1 \ldots Y_n \overset{iid}{\sim} \text{Exp}(\theta)$. Find an unbiased estimate for $Y = X_{(1)}$, the sample min.

  - $f(x) = \frac{1}{\theta} e^{-x/\theta}$, $x \geq 0$, $\theta > 0$

  - $F(x) = 1 - e^{-x/\theta}$

  - $f_{Y_{(1)}}(x) = \frac{n!}{(n-1)!} \frac{1}{\theta} e^{-x/\theta} [1 - (1 - e^{-x/\theta})]^{n-1} = \frac{n}{\theta} e^{-x/\theta} e^{-x(n-1)/\theta} = \frac{n}{\theta} e^{-xn/\theta}$ So $f(y) = \frac{n}{\theta} e^{-yn/\theta}$

- Then Bias = $E(Y) - \theta$, so $E(Y) = \int y \frac{n}{\theta} e^{-yn/\theta}\,dy = \frac{n}{\theta} \int y e^{-yn/\theta}\,dy$ Let $u = y$, $du = dy$, $dv = e^{-yn/\theta}$, $v = \int e^{-yn/\theta}\,dy = -\frac{\theta}{n} e^{-yn/\theta}$

$= \left(\frac{n}{\theta}\right)\left(\frac{\theta}{n}\right)^2 = \frac{\theta}{n}$ or, $\int y e^{-yn/\theta}\,dy \sim$

$\text{Ga}(2, n/\theta)$ so integrates to $\Gamma(2)(n/\theta)^{-2} = \left(\frac{\theta}{n}\right)^2$

- Estimator is biased: $E(\hat\theta) - \theta = \frac{\theta}{n} - \theta \neq 0$

- To find unbiased: $a\left(\frac{\theta}{n}\right) - \theta = 0 \longrightarrow a = \theta\left(\frac{n}{\theta}\right) \longrightarrow a = n$, so

use $\hat\theta = nY = nX_{(1)}$ as the unbiased estimator.

· Ex: $z_i = \begin{cases} \text{"blue" w. prob } \theta \\ \text{"red" w. prob } 1-\theta \end{cases}$ $\quad x_i \sim \begin{cases} \text{Pois}(1) \text{ if } z_i = \text{"blue"} \\ \text{Pois}(4) \text{ if } z_i = \text{"red"} \end{cases}$

Only observe x_i's. Find consistent estimator of θ. $\quad (\hat{\theta} \xrightarrow{p} \theta)$

· $E(x_i) = E(E(x_i|z)) = 1\theta + 4(1-\theta) = 4-3\theta$, so $E(x_i) = 4-3\theta$

So since $\bar{x} \xrightarrow{p} E(x)$ by the WLLN, use $\bar{x} = 4-3\theta \rightarrow \hat{\theta} = \frac{4-\bar{x}}{3}$,

So by the WLLN, $\hat{\theta} \xrightarrow{p} \theta$.

Max Likelihood Estimation Review:

· Binom(n,p) example: $x_1 \dots x_n \overset{iid}{\sim} \text{Bin}(n,p)$. Find mle of p.

· Lhood $= \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \propto p^{\Sigma x_i}(1-p)^{\Sigma n - x_i}$? Nope - we always add binomial RVs together to get 1 Binomial RV, with n = sum of n's.

So: · Lhood for $x \sim \text{Binom}(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$, ie $\propto p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$ for x_i = Bernoulli trials.

$$\log L \propto \sum_{i=1}^{n} x_i \log p + (n - \Sigma x_i)\log(1-p)$$

$$\frac{\partial}{\partial p}\log L = \frac{\Sigma x_i}{p} - \frac{(n-\Sigma x_i)}{1-p} = 0 \rightarrow \frac{\hat{p}}{1-\hat{p}} = \frac{\Sigma x_i}{n - \Sigma x_i} \rightarrow \frac{\hat{p}}{1-\hat{p}} = \frac{\bar{x}}{1-\bar{x}} \quad \text{so } \hat{p} = \bar{x}$$

Jeffreys Prior

· A vague prior, which can lead to a proper or improper prior.

· Equal to sqrt(Fisher's info)

· Jeffreys Prior $\propto \sqrt{\det I(\theta)}$

1) Univariate case: $\pi(\theta) \propto \sqrt{I(\theta)}$, $\quad I = -E_x\left[\frac{-d^2}{d^2\theta}\log L\right]$

2) Multivariate case: $\pi(\theta) \propto \sqrt{\det(I(\theta))}$, $\quad \underset{p\times p}{I} = -E_x\left[\left\{\frac{d^2}{d\theta_i d\theta_j}\log L\right\}_{ij}\right]$

· Ex 1: $X \sim \text{Binom}(n,p)$ for a single obs x, with n known. $\log L \propto x\log p + (n-x)\log(1-p)$

$$I(p) = -E_x\left[\frac{d^2}{d^2p}\log L\right] \Rightarrow \frac{d\log L}{dp} = \frac{x}{p} - \frac{(n-x)}{(1-p)} \rightarrow \frac{d^2\log L}{dp^2} = \frac{x}{p^2} + \frac{(n-x)}{(1-p)^2}$$

$$\text{so } I(p) = -E_x\left[\frac{x}{p^2} + \frac{n-x}{(1-p)^2}\right] \quad \text{and } E(x) = np, \text{ so } = -\left[\frac{np}{p^2} + \frac{n-np}{(1-p)^2}\right] = \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}$$

Then Jeffreys $\pi(\theta) \propto \sqrt{\frac{n}{p(1-p)}} \propto p^{-1/2}(1-p)^{-1/2} \sim \text{Beta}(1/2, 1/2)$

· Ex 2: $x|\theta \sim N(\mu, \sigma^2)$ so $\theta = (\mu, \sigma^2)$. Again, let X be a single obs.

· $\log L \propto -\log(\sigma) - \frac{1}{2\sigma^2}(x-\mu)^2$.

· 2nd partial derivs:

(1st) → · $\frac{\partial}{\partial\mu}\log L = +\frac{2(x-\mu)}{2\sigma^2} = \frac{(x-\mu)}{\sigma^2}$ · $\frac{\partial \log L}{\partial\sigma} = \frac{-1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$

(2nd) → · $\frac{\partial^2 \log L}{\partial\mu^2} = \frac{-1}{\sigma^2}$ · $\frac{\partial^2 \log L}{\partial\mu\partial\sigma} = \frac{\partial^2 \log L}{\partial\sigma\partial\mu} = \frac{-2(x-\mu)}{\sigma^3}$ · $\frac{\partial^2 \log L}{\partial\sigma^2} = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$

· Expected vals: · $E_x\left(\frac{\partial^2 \log L}{\partial\mu^2}\right) = -1/\sigma^2$ · $E_x\left(\frac{\partial^2 \log L}{\partial\mu\partial\sigma}\right) = 0$ · $E_x\left(\frac{\partial^2 \log L}{\partial\sigma^2}\right) = -2/\sigma^2$

· Fishers m/o: $I(\theta) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix}$

· Jeffreys: $\pi(\theta) = \sqrt{\det I(\theta)} = \sqrt{(1/\sigma^2)(2/\sigma^2) - 0} = \sqrt{2/\sigma^4} = \sqrt{2}/\sigma^2 \propto 1/\sigma^2$, so

$\pi(\theta) \propto \frac{1}{\sigma^2}$ for $\sigma > 0$.

Deriving Densities of Order Stats $X_{(1)} \sim X_{(n)}$:

- For any order stat, pdf is: $f(x_{(j)}) = \frac{n!}{(j-1)!(n-j)!} f(x) F(x)^{j-1} [1-F(x)]^{n-j}$

- Can easily derive the CDFs for $X_{(1)} \sim X_{(n)}$: Say $X_1 \dots X_n \overset{iid}{\sim} F(x)$.

1) $Y = X_{(1)}$, sample min: $P(Y \le y) = P(\min(X_1 \dots X_n) \le y) = 1 - P(\text{all } X_1 \dots X_n \ge y)$

- Then since X's iid, $= 1 - (P(X \ge y))^n = 1 - [1 - P(X \le y)]^n$ is cdf $F(y)$. $= 1 - [1 - F_x(y)]^n$

- For pdf, $f(y) = \frac{\partial F(y)}{\partial y} = -n[1-F_x(y)]^{n-1}(-f_x(y)) = n[1-F_x(y)]^{n-1} f_x(y)$

2) $Y = X_{(n)}$, sample max: $P(Y \le y) = P(\text{all } x_i \le y) = [P(X \le y)]^n = F_x(y)^n$

Then pdf is $n F_x(y)^{n-1} f_x(y)$

Transformations Review: Non-monotone:

- $f(x) = \frac{1}{2} e^{-|x|}$, $-\infty < x < \infty$. $Y = |X|^3$. Find $f(y)$.

- $A_1 = (0, \infty)$: $f(x) = \frac{1}{2} e^{-x} \rightarrow g(x) = x^3$, so $g^{-1}(y) = y^{1/3}$
- $A_2 = (-\infty, 0)$: $f(x) = \frac{1}{2} e^{x} \rightarrow g(x) = x^3$, so $g^{-1}(y) = -y^{1/3}$

} y has same support on both, $y \ge 0$

- $|J|$ · for A_1: $\left|\frac{\partial}{\partial y} g^{-1}(y)\right| = \frac{1}{3} y^{-2/3}$ · for A_2: $\left|\frac{\partial}{\partial y} g^{-1}(y)\right| = -\frac{1}{3} y^{-2/3} = \frac{1}{3} y^{-2/3}$

- $f_Y(y) = \frac{1}{2} e^{-(y^{1/3})} \left(\frac{1}{3} y^{-2/3}\right) + \frac{1}{2} e^{-(y^{1/3})} \left(\frac{1}{3} y^{-2/3}\right) = \frac{1}{3} y^{-2/3} e^{-y^{1/3}}$ for $y > 0$.

Mgfs of Transformations:

- For $X \sim \exp(\beta)$, we can find mgf: $M_X(t) = E(e^{Xt}) = \int_0^\infty e^{xt} \frac{1}{\beta} e^{-x/\beta} dx = \int_0^\infty \frac{1}{\beta} e^{-x(1/\beta - t)} dx$

Let $u = x(\frac{1}{\beta} + t)$ or recognize $\exp(1/\beta - t)$ kernel, so int to $1/(1/\beta - t)$.

$= \frac{1}{\beta}\left(\frac{1}{1/\beta - t}\right) = \frac{1}{1-\beta t}$

- Say $X \sim ga(a, \beta)$ with $M_X(t) = \left(\frac{1}{1-\beta t}\right)^a$. Find dist of $W = \frac{2X}{\beta}$ using mgfs.

$M_W(t) = E\left[e^{\frac{2X}{\beta} t}\right] = \dots$ (solve the usual way now), $\int e^{2xt/\beta} f(x) dx$

- Ex: Sum of two exponentials: $X_1 \perp\!\!\!\perp X_2 \sim Exp(\beta)$. Find dist of $W = X_1 + X_2$ using mgfs.

$M_W(t) = M_{X_1+X_2}(t) = E[e^{(X_1+X_2)t}] = E(e^{X_1 t} e^{X_2 t}) = M_{X_1}(t) M_{X_2}(t) = \left(\frac{1}{1-\beta t}\right)^2 \sim ga(2, \beta)$

Joint Distributions:

- Consider $f(y_1, y_2) = 4 y_1 y_2$ for $0 \le y_1 \le 1$, $0 \le y_2 \le 1$.

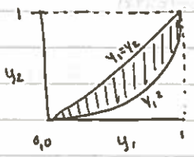
- $P(y_1 \le 1/2, y_2 \le 3/4) = \int_0^{3/4}\int_0^{1/2} 4 y_1 y_2 \, dy_1 dy_2 = \int_0^{3/4} 4\frac{y_1^2}{2}\Big|_{y_1=0}^{1/2} y_2 dy_2 = \int_0^{3/4} \frac{1}{2} y_2 dy_2 = \frac{1}{2}\frac{y_2^2}{2}\Big|_{y_2=0}^{3/4} = \frac{1}{4}\left(\frac{3}{4}\right)^2 = \frac{9}{64}$

- $P(y_1 \le 1/2 \mid y_2 \le 3/4) = \frac{P(y_1 \le 1/2, y_2 \le 3/4)}{P(y_2 \le 3/4)}$ Numerator is $9/64$.

- $P(y_2 \le 3/4) = \int_0^{3/4}\int_0^{1} 4 y_1 y_2 \, dy_1 dy_2 = \int_0^{3/4} 4 y_2 \frac{y_1^2}{2}\Big|_0^1 dy_2 = \int_0^{3/4} \frac{4}{2} y_2 dy_2 = y_2^2\Big|_0^{3/4} = 3/4$

- $\frac{9/64}{3/4} = \frac{9}{64} \cdot \frac{4}{3} = \frac{12}{64} = \frac{3}{16}$

- $P(y_1^2 < y_2 < y_1) = \int_0^1 \int_{y_1^2}^{y_1} 4 y_1 y_2 \, dy_2 y_1 = 1/6$



Bivariate Transformations

1. Check that the two forms being transformed are 1-1 & onto.
   Say considering a transform from $(x,y) \to (u,v)$.
   - (1-1): Let $u_1, v_1 = u_2, v_2$, show implies $x_1, y_1 = x_2, y_2$. (ie $x_1 = x_2$ and $y_1 = y_2$)
   - Onto: (u,v) yields a unique (x,y).

2. $|J| = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}$ where h_1 & h_2 are the $g^{-1}(y)$

3. $f(u,v) = f(h_1, h_2)|J|$ where if vars indep, $= f(h_1) f(h_2) |J|$.

Ex 1: Product of Two Betas. Let $X \sim Beta(a,b)$, $Y \sim Beta(a+b, \gamma)$. $X \perp\!\!\!\perp Y$.
Find pdf of XY. Note: $0<x<1$, $0<y<1$.

- Let $u = xy$, $v = y$. Then $v = y$, and $x = u/v$. $h_1 = u/v$, $h_2 = v$
- Both forms are 1-1 & onto:
- Jacobian: $|J| = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix} = \begin{vmatrix} 1/v & -u/v^2 \\ 0 & 1 \end{vmatrix} = 1/v$
- pdf: $f_{uv}(u,v) = \frac{1}{B(a,b)} (u/v)^{a-1} (1-u/v)^{b-1} \cdot \frac{1}{B(a+b,\gamma)} v^{a+b-1} (1-v)^{\gamma-1} v^{-1}$

$$= \frac{\Gamma(a+b+\gamma)}{\Gamma(a)\Gamma(b)\Gamma(\gamma)} (u/v)^{a-1} (1-u/v)^{b-1} v^{a+b-1} (1-v)^{\gamma-1} v^{-1}$$

- $f_u = \int_u^1 \frac{\Gamma(a+b+\gamma)}{\Gamma(a)\Gamma(b)\Gamma(\gamma)} u^{a-1} v^{-a} v^{a+b-1} (1-u/v)^{b-1} (1-v)^{\gamma-1} dv$ and so forth.

(since $x = u/v$ in $(0,1)$)

Ex 2. $X \perp\!\!\!\perp Y \sim$ gamma, with $X \sim ga(r,1)$, $Y \sim ga(s,1)$. Let $z_1 = x+y$, $z_2 = \frac{x}{x+y}$.
Find dists of z_1, z_2.

- Check 1-1 & onto ✓
- $z_1 = x+y \to x = z_1 - y \to z_2 = \frac{z_1 - y}{z_1 - y + y} \to z_1 z_2 = z_1 - y \to y = z_1 - z_1 z_2$

  Then $z_1 = x + z_1 - z_1 z_2 \to x = z_1 z_2$. So $h_1(z_1, z_2) = z_1 z_2$, $h_2(z_1, z_2) = z_1 - z_1 z_2$
- $|J| = \begin{vmatrix} \frac{\partial h_1}{\partial z_1} & \frac{\partial h_1}{\partial z_2} \\ \frac{\partial h_2}{\partial z_1} & \frac{\partial h_2}{\partial z_2} \end{vmatrix} = \begin{vmatrix} z_2 & z_1 \\ (1-z_2) & -z_1 \end{vmatrix} = \left| -z_1 z_2 - z_1(1-z_2) \right| = z_1$
- pdf: $f(z_1, z_2) = \frac{1}{\Gamma(r)} (z_1 z_2)^{r-1} e^{-z_1 z_2} \cdot \frac{1}{\Gamma(s)} (z_1 - z_1 z_2)^{s-1} e^{-z_1 + z_1 z_2} \cdot z_1$

$$= \frac{1}{\Gamma(r)\Gamma(s)} z_1^r z_2^{r-1} e^{-z_1} z_1^{s-1} (1-z_2)^{s-1} = \underbrace{\frac{1}{\Gamma(r)\Gamma(s)} z_1^{r+s-1} e^{-z_1}}_{z_1 \sim ga(r+s,1)} \cdot \underbrace{z_2^{r-1} (1-z_2)^{s-1}}_{z_2 \sim Beta(r,s).}$$

- Convolutions: If working with the transformation $Z = X+Y$,

$$f(z) = \int_{-\infty}^{\infty} f_x(z-y) f_y(y) dy \quad \text{where } f_x, f_y \text{ are pdfs for } x, y.$$

- Ex: Sum of two exponentials w. same parameter.
  $X \sim Exp(\lambda)$, $Y \sim Exp(\lambda)$, $X \perp\!\!\!\perp Y$. Let $Z = X+Y$. Find pdf of Z.
- $f(x) = \lambda e^{-\lambda x}$, $f(y) = \lambda e^{-\lambda y}$, $f(z) = \int_0^z \underbrace{\lambda e^{-\lambda(z-y)}}_{f_x(z-y)} \cdot \underbrace{\lambda e^{-\lambda y}}_{f_y(y)} dy$

$$= \int_0^z \lambda^2 e^{-\lambda z} dy = z\lambda^2 e^{-\lambda z} \Big|_{z=0}^{z} = \lambda^2 z e^{-\lambda z} \propto z e^{-\lambda z} \sim ga(2, \lambda)$$

Math Stats Flashcard Facts:

1. Delta Method: If $Y_n \sim N(\theta, \sigma^2/n)$ for $Y_n = g(X_n)$, then can approx the mean + var of Y as:

$$\sqrt{n}\,(g(X_n) - g(\theta)) \sim N(0, \sigma^2 [g'(\theta)]^2) \quad \text{i.e. } g(X_n) \xrightarrow{d} N\left(g(\theta), \frac{\sigma^2}{n}(g'(\theta))^2\right)$$

for $g'(\theta)$ differentiable, ie $\neq 0$. Can use $n=1$ for a single RV.

- Why it works: can write the 2nd order Taylor series of $g(X_n)$ around θ:

$$g(X_n) \approx g(\theta) + g'(\theta)(X_n - \theta) + \varepsilon \qquad \varepsilon \to 0 \text{ as } n \to \infty$$

So $\sqrt{n}\,(g(X_n) - g(\theta)) \approx g'(\theta)(X_n - \theta) \sim N(0, \text{Var}(X_n) \cdot g'(\theta)^2)$

- Ex: $X \sim U(0,1)$. $Y = \sqrt{X}$. Approximate $E(Y)$ + $\text{Var}(Y)$. - $E(X) = 1/2$, $\text{Var}(X) = 1/12$.

- $E(Y) \approx \sqrt{1/2}$
- $\text{Var}(Y) \approx \frac{1}{12} \cdot \left[\frac{1}{2(1/2)^{1/2}}\right]^2 = \frac{1}{24}$
- $g'(x) = \frac{\partial}{\partial x} x^{1/2} = \frac{1}{2x^{1/2}}$

2. Formula for Taylor Series: of $f(x)$ around θ: $f(x) = f(\theta) + \frac{f'(\theta)(x-\theta)}{1!} + \frac{f''(\theta)(x-\theta)^2}{2!} + \dots + e$

3. Limit definition of e: $\lim_{x \to \infty} \left(1 + \frac{a}{x}\right)^{bx} = e^{ab}$

4. e^x in summation form: $\sum_{k=0}^{\infty} x^k/k! = e^x$

5. Infinite geometric Series: $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$

6. Binomial Summation: $\sum_{k=0}^{\infty} \binom{k}{k} x^{n-k} y^k = (x+y)^n$

7. Gamma fcn: $\Gamma(x) = (x-1)!$

8. Binomial Coefficient: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

3. Inequalities ~ Convergence:

- Markov's Inequality: $P(g(x) \geq r) \leq \frac{E(g(x))}{r}$

- Chebyshev's Inequality: $P(|x-\mu| > t) \leq \frac{\sigma^2}{t^2}$ or $P(|x-\mu| \geq \sigma t) \leq \frac{1}{t^2}$, $\mu = E(x)$, $\sigma^2 = \text{Var}(x)$.

- Derive from Markov's: Let $t^2 = r$, $g(x) = (x-\mu)^2/\sigma^2$. Then

$$P(g(x) \geq r) \leq \frac{E(g(x))}{r} \to P\left(\frac{(x-\mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{E[(x-\mu)^2]}{\sigma^2 t^2}$$

$$\to P(|x-\mu| \geq t\sigma) \leq \frac{\sigma^2}{\sigma^2}\frac{1}{t^2} \quad \text{by multiplying LHS in } P(\cdot) \text{ by } \sigma^2, \text{ and taking sqrt.}$$

- Slutsky's Lemma: Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, c is constant. Then
  - $X_n + Y_n \xrightarrow{d} X + c$
  - $X_n Y_n \xrightarrow{d} Xc$

- Cauchy - Schwartz Inequality:
  - $(x'y)^2 \leq (x'x)(y'y)$ also $E(xy)^2 \leq E(x^2)\,E(y^2)$

- Stein's Lemma: $\text{cov}(g(x), Y) = E(g'(x))\,\text{cov}(x, y)$ when x, Y are jointly normal.

- Central Limit Theorem $\bar{X} \xrightarrow{d} N(\mu, \sigma^2/n)$ for $\mu = E(x)$, $\sigma^2 = \text{Var}(x)$

- WLLN: $\bar{X}_n \xrightarrow{p} \mu$, ie $\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$

- SLLN: $\bar{X}_n \xrightarrow{a.s.} \mu$, a.s. = almost surely

- Convergence in Prob: $X_1 \dots X_n \xrightarrow{p} X$, ie $X_n \xrightarrow{p} X$, iff

$$\lim_{n \to \infty} P(|X_n - X| \geq \varepsilon) = 0 \equiv \lim_{n \to \infty} P(|X_n - X| < \varepsilon) = 1$$

- Convergence in Dist: $X_n \xrightarrow{D} X$ iff $\lim_{n \to \infty} F_{X_n}(x) = F(x)$

- Order of Strength of Convergence:

distribution
probabilities
means
almost surely
surely

(weak → strong)

4. Useful Definitions + Tricks:

- Chain Rule of Probability: $p(a,b,c,d) = p(a)\,p(b|a)\,p(c|b,a)\,p(d|c,b,a)$

- Cov + Sample Cov: $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)^T] = E(XY^T) - E(X)E(Y)^T$

$$\hat{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- Conditional Independence: If $X \perp\!\!\!\perp Y \mid Z$, then
  - $P(X|Z)P(Y|Z) = P(X,Y|Z)$
  - $P(X|Y,Z) = P(X|Z)$

- Conditional Expectation + Variance: For ANY X and Y,
  - $E(X) = E(E(X|Y))$
  - $Var(X) = E(Var(X|Y)) + Var(E(X|Y))$ · "EVVE"

- This also holds if conditioning on something else, ie $E(X|Y) = E(E(X|Y,Z))$

- Covariance is a good place to use this:

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X\cdot\underline{E(Y|X)}) - E(X)\,E(\underline{E(Y|X)})$$

- Easiest way to find $E(1/X)$: No transform, just $\int \frac{1}{x} f(x)\,dx$.

- Independence: If $X \perp\!\!\!\perp Y$, then $g(x) \perp\!\!\!\perp h(y)$.

- Switching order of sums + integrals: If $h(x,\theta)$ continuous in θ, and $\sum_{x=0}^{\infty} h(x,\theta)$ converges uniformly on $[a,b]$, then

$$\int_a^b \sum_{x=0}^{\infty} h(x,\theta)\,d\theta = \sum_{x=0}^{\infty}\int_a^b h(\theta,x)\,d\theta$$

5. Finding Sufficient Statistics:

- If can factor Lhood as $f(x|\theta) = g(T(x)|\theta)\cdot h(x)$, then $T(x)$ is a suff stat.

- To find a minimal suff stat, take $f(x|\theta)/f(y|\theta)$, ratio of Lhoods. The suff stat is term which makes ratio constant wrt θ. (Doesn't have to be 1)

6. Distributional Relationships:

- Normal + Chisq:
  - $Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$
  - $Z_1 \ldots Z_n \overset{iid}{\sim} N(0,1)$ then $Z_1^2 + \ldots + Z_n^2 \sim \chi^2_n$

- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

- $F = \frac{U/u}{V/v}$ where $U \sim \chi^2_u$ + $V \sim \chi^2_v$. Then $1/F \sim F_{v,u}$

- Exp(λ) cdf: $F(x) = 1 - e^{-\lambda x}$

- Exp(λ) to Pois(tλ): If $t \sim Exp(\lambda)$ is time btwn successive arrivals, let N be # of arrivals in a fixed period t. Then $N \sim Pois(t\lambda)$

- Sum of Exponentials + Gammas:
  - Exp(1) is $Ga(1,\lambda)$
  - Sum of $X_1 \ldots X_n \overset{iid}{\sim} Ga(\alpha_i, b) = Ga(\Sigma\alpha_i, b)$
  - Sum of $X_1 \ldots X_n \overset{iid}{\sim} Exp(\lambda) = Ga(n,\lambda)$

- Distribution of an order stat $X_{(j)}$: $f(x_{(j)}) = \frac{n!}{(j-1)!(n-j)!} f(x)\,F(x)^{j-1}[1-F(x)]^{n-j}$

- t and F relationship: If $X \sim t_n$, $X^2 \sim F(1,n)$

- Uniform cdf for $U(a,b)$: $\frac{x-a}{b-a}$

- Sum of Indep Poissons: $X \sim Pois(\lambda)$, $\perp\!\!\!\perp$ $Y \sim Pois(\theta)$, then $X+Y \sim Pois(\lambda+\theta)$

- Ratio of 2 std normals $X/Y \sim Cauchy(0,1)$

- Sum of Chi-Squareds: χ^2, with df = sum of dfs

- $T = Z/\sqrt{U/u}$ for $T \sim t_n$, $Z \sim N(0,1)$, $U \sim \chi^2_u$

- Dist of $\frac{(\bar{X}-\mu)}{S/\sqrt{n}} \sim t_{n-1}$

- Inverse Gamma mean + cov: $X \sim IG(a,b)$. $E(X) = \frac{b}{a-1}$ $Var(X) = \frac{b^2}{(a-1)^2(a-2)}$

7. MGFs

- Thm 2.3.11 - Moments defining cdfs.
  - If X, Y have bounded support, then $X=Y$ iff $E(X^r)=E(Y^r)$, ie all moments same.
  - If mgfs exist + are equal in h neighborhood of $t=0$, mgfs are same.

- If $X \perp\!\!\!\perp Y$, $Z=X+Y$, then $M_Z(t) = M_X(t)\, M_Y(t)$

- Mgf of $aX+b$: $E(e^{(ax+b)t}) = e^{bt} E(e^{Xat}) = e^{bt} M_X(at)$

- Mgf of $\bar{X}$: $M_{\bar{X}}(t) = \left[M_X(t/n)\right]^n$

8. Exponential families: $f(x|\theta) = h(x) \cdot c(\theta) \exp\left[\sum_{i=1}^{k} w_i(\theta) t_i(x)\right]$
  - If $k = \text{length}(\theta)$, is full exp family.
  - If $k < \text{length}(\theta)$, is curved exp family.

- $T(x) = \left[\sum_{i=1}^{n} t_1(x_i), \ldots, \sum_{i=1}^{n} t_k(x_i)\right]$ is a sufficient statistic.

- If full exp family, $T(x)$ is a complete sufficient statistic.

- NOT exp families: t, F, Cauchy, Binom if n unknown, NBinom if r unknown, Weibull if shape unknown.