STA290

Winter 2014

Selected presentation materials

Contents

Chapter 1

Abstract vector spaces

1.1 Coordinate free projections

In this chapter we consider some abstract vector space, call it V, which has a norm $\|\cdot\|$ induced by an inner product $\langle\cdot,\cdot\rangle$ (i.e. $\|v\|^2 := \langle v, v\rangle\rangle$). The elements of V are called vectors but since we are considering abstract spaces we do not necessarily have access to the "coordinates" of each vector. What we do have is the ability to construct an orthonormal basis of V, usually denoted something like $\phi_1, \phi_2, \ldots, \phi_d$, where d is the dimension of V. Then each vector in v has a decomposition in terms of these basis vectors. This note explores to computing things like norms and inner products and taking projections with these basis vectors.

The idea is that V will denote something like points in space. In physics, points in space should exist independently of what coordinate system one uses. In particular, the notion of distance and inner product exist *b*efore we attach a coordinate system to this space. The following chapter talks about how to work with such vectors. The main story is that one can choose a particular orthonormal basis and then the coefficients of the basis decomposition behave exactly as regular coordinates in $\mathbb{R}^d$. Moreover, if one wants to project a vector in V to a linear subspace $M \subset V$ then by choosing the basis vectors appropriately one can easily perform the desired projection by simply truncating the basis representation.

Claim 1 (Coefficients are coordinates). *Let $\phi_1, \ldots, \phi_d$ be an orthonormal basis for V. Then any $v \in V$ has a unique representation*

$$\boxed{v = \sum_{k=1}^{d} c_k \phi_k, \text{ where } c_k = \langle v, \phi_k \rangle.}$$

Moreover, these basis coefficients behave exactly like coordinates in regular Euclidean space so that if $v = \sum_{k=1}^{n} c_k \phi_k$ *and* $w = \sum_{k=1}^{n} d_k \phi_k$*, then*

$$\boxed{\langle v, w \rangle = \sum_{k=1}^{d} c_k d_k.}$$

In particular, $\|v\|^2 = \sum_{k=1}^{d} c_k^2$.

Now suppose $M \subset V$ is an m-dimensional linear subspace (the zero vector should be in here). We can define $M^\perp$ to be the set of all vectors in V which are orthogonal to every vector in M. In this case we can write $M \oplus M^\perp = V$ to signify that every $v \in V$ has a unique decomposition $v_M + v_M^\perp$ where $v_M \in M$ and $v_M^\perp \in M^\perp$. Often, in statistics, one needs to project a vector $v \in V$ to a subspace M. This operation is denoted $P_M v$ and is technically defined as $P_M v := \text{argmin}_{w \in M} ||w - v||^2$. The following theorem shows that with a judicious choice of your orthonormal basis this projection is easily calculated

Claim 2 (Projections are easy). *If one constructs the orthonormal basis* $\phi_1, \ldots, \phi_d$ *of* V *in such a way that*

$$M = span\{\phi_1, ..., \phi_m\}$$

then for any vector $v = \sum_{k=1}^{d} c_k \phi_k$ *the projection to* M *is computed by truncating the decomposition to* m*:*

$$\boxed{P_M v = \sum_{k=1}^{m} c_k \phi_k}. \tag{1.1}$$

Moreover, since $P_M v$ *must be orthogonal to* $P_{M^\perp} v$*, and* $P_{M^\perp} v = v - P_M v$*, we have that*

$$\boxed{P_M v \perp (v - P_M v).}$$

The following is a simple consequence of the above claim, but it will be important later so we state it here

Corollary 1. *If* M *is a linear subspace of* V *and* $v, w \in V$ *then*

$$\langle v, P_M w \rangle = \langle P_M v, P_M w \rangle = \langle P_M v, w \rangle.$$

1.2 Projections for Gaussians

If we are working in an abstract vector space V, what do we even mean by a Gaussian vector in V? A random vector Y is said to be Gaussian if $\langle v, Y\rangle$ is a Gaussian random variable for all $v \in V$. Now we want some notion of $Y \sim \mathcal{N}(0, \sigma^2 I_d)$. To be able to say $Y \sim \mathcal{N}(0, \sigma^2 I_d)$ we simply require that there exists a orthonormal basis representation $Y = \sum_{k=1}^{d} d_k \psi_k$ where $d_1, \ldots, d_d$ are iid $\mathcal{N}(0, \sigma^2)$.

A fundamental fact about independent mean zero Gaussian random variables is that they are invariant under rotations. In particular if $W \sim \mathcal{N}\left(0, \sigma^2 I_d\right)$ where W is a random vector in Euclidean space then $UW \sim \mathcal{N}\left(0, \sigma^2 I_d\right)$ for any orthogonal rotation matrix (U is a rotation if $I = U^t U$). In fact, independent mean zero Gaussians make up the only multivariate distribution with independent coordinates that is rotationally invariant. You can think of left multiplication by a rotation matrix as simply changing basis. Therefore if Y is an abstract random vector that satisfies $Y \sim \mathcal{N}(0, \sigma^2 I)$, then *any* basis decomposition of Y should give iid $\mathcal{N}(0, \sigma^2)$ coefficients.

Claim 3. *Suppose Y is an abstract Gaussian random vector in V which satisfies $Y \sim \mathcal{N}(0, \sigma^2 I)$. If $\phi_1, \ldots, \phi_d$ is a orthonormal basis of V then*

$$\boxed{Y = \sum_{k=1}^{d} c_k \phi_k \text{ implies } c_1, \ldots, c_d \text{ are iid } \mathcal{N}(0, \sigma^2).}$$

Once we have the above theorem the following corollary is easy to prove.

Corollary 2. *Suppose $\phi_1, \ldots, \phi_d$ is a orthonormal basis of V and $Y \sim \mathcal{N}(0, \sigma^2 I_d)$. Suppose M is an m-dimensional linear subspace of V which is spanned by $\phi_1, \ldots, \phi_m$. If $Y = \sum_{k=1}^{d} c_k \phi_k$ and $v = \sum_{k=1}^{d} v_k \phi_k$ then the following three statements are true:*

(i) $||P_M Y||^2 = c_1^2 + \cdots c_m^2 \sim \sigma^2 \chi_m^2$;

(ii) $P_M Y$ *is independent of* $Y - P_M Y$;

(iii) $var(\langle v, Y\rangle) = var(v_1 c_1 + \cdots v_d c_d) = \sigma^2 \|v\|^2$.

Just as a simple example of power and simplicity of these statements notice that one can easily recover the fact that $\overline{Y}$ and S^2 are independent.

1.3 Application to linear models

The basic coordinate-free linear model

$$Y = \mu + Z$$

where Y, μ and Z all take values in a vector space V, the mean vector μ is assumed to be in a linear subspace $M \subset V$ and $Z \sim \mathcal{N}(0, \sigma^2 I)$.

1.3.1 Gauss-Markov

The Gauss-Markov theorem effectively says that the optimal estimate of μ is the projection of Y onto M, i.e. $\hat{\mu} = P_M Y$.

Claim 4 (Gauss-Markov)**.** *For any $v \in M$, the minimum variance linear unbiased estimate of $\langle v, \mu \rangle$ is given by $\langle v, \hat{\mu} \rangle$ where $\hat{\mu} := P_M Y$.*

Proof. There are two main facts in the proof of the Gauss-Markov Theorem. First, notice

$$\langle v, \hat{\mu} \rangle = \langle v, P_M Y \rangle = \langle P_M v, P_M Y \rangle = \langle P_M v, Y \rangle \tag{1.2}$$

which follows by Corollary 1. The second fact is to simply unravel what it means for a linear estimate $\langle w, Y \rangle$ to be unbiased. Notice that $E\langle w, Y \rangle = \langle w, \mu \rangle = \langle P_M w, \mu \rangle$ so the unbiased constraint becomes

$$\langle P_M w, \, \mu \rangle = \langle P_M v, \, \mu \rangle \text{ for all } \mu \in M. \tag{1.3}$$

Now this implies that $P_M w = P_M v$ (because the above formula implies $\langle P_M w - P_M v, \nu \rangle = 0$ for $\nu := P_M w - P_M v \in M$). To summarize

> *A necessary and sufficient condition that $\langle w, Y \rangle$ be a linear unbiased estimate of $\langle v, \mu \rangle$ is that $P_M w = P_M v$.*

By (1.2) we know that the Gauss-Markov estimate is a linear unbiased estimate. Now it is easy to show the Gauss-Markov estimator has minimum variance among linear unbiased estimates

$$\begin{aligned}
\operatorname{var}(\langle w, Y \rangle) &= \operatorname{var}(\langle w, Z \rangle) \\
&= \sigma^2 \|w\|^2, \quad \text{by 2} \\
&= \sigma^2 \|P_M w\|^2 + \sigma^2 \|P_M^{\perp} w\|^2 \\
&\geq \sigma^2 \|P_M w\|^2 \\
&= \sigma^2 \|P_M v\|^2, \quad \text{if } \langle w, Y \rangle \text{ is unbiased} \\
&= \operatorname{var}(\langle P_M v, Y \rangle) \\
&= \operatorname{var}(\langle v, \hat{\mu} \rangle).
\end{aligned}$$

□

1.3.2 Regression

Notice that the Gauss-Markov theorem has implications for ANOVA, regression and most other linear statistical models. In this section we explore the Gauss-Markov theorem in the context of regression. The standard regression setup is

$$Y = X\beta + Z$$

where X is the design matrix with p columns and β is a p dimensional coefficient vector of unknowns and $Z \sim \mathcal{N}(0, \sigma^2 I)$. This is indeed equivalent to our basic linear model

$$Y = \mu + Z$$

where now μ is assumed to be in the linear space $M := \{\beta_1 x_1 + \cdots \beta_p x_p : \beta_k \in \mathbb{R}\}$ where x_k is the k^{th} column of X.

To reconcile the OLS estimate $\hat{\beta} = (X^t X)^{-1} X^t Y$ and the Gauss-Markov estimate $\hat{\mu} = P_M Y$ simply notice that the OLS estimate of μ is simply $X\hat{\beta} = X(X^t X)^{-1} X^t Y$ which is the same as $P_M Y$. In particular, $X(X^t X)^{-1} X^t$ is the coordinate-matrix form of P_M. This is easy to see when one adds the additional assumption that the columns of X are orthonormal. In this case we know $X^t X = I$ and from Claim 2 we know how to do the projection:

$$\begin{aligned} P_M Y &= \sum_{k=1}^{p} \langle Y, x_k \rangle x_k \\ &= X X^t Y, \quad \text{since } X^t Y \text{ gives } \langle Y, x_k \rangle \\ &= X(X^t X)^{-1} X^t Y \\ &= X\hat{\beta}. \end{aligned}$$

Now a few nice facts from regression automatically follow from the basic facts of projection. In particular we know that P_M is independent of $Y - P_M Y$ from Corollary 2. This translates to the fact that

> *The estimated residuals $Y - X\hat{\beta}$ and the estimated mean $X\hat{\beta}$ are independent (and orthogonal as vectors).*

We also know that when μ is the zero vector (i.e. there is no signal) then $\|\hat{\mu}\|^2 = \|X\hat{\beta}\|^2 = \|P_M Y\|^2 \sim \sigma^2 \chi^2_p$. In a similar manner it is also clear that $\|Y - \hat{\mu}\|^2 \sim \sigma^2 \chi^2_{n-p}$. This gives the classic F test for $H_0 : \mu = 0$

Under the null hypothesis that $\mu = 0$ the random variable $F := \|\hat{\mu}\|^2/\|Y - \hat{\mu}\|^2$ has the same distribution as Z_1/Z_2 where $Z_1 \sim \chi^2_p$ and $Z_2 \sim \chi^2_{n-p}$ are independent random variables.

Note that the independence again comes from the fact that projections of spherical Gaussians are independent of their residuals. This way of viewing the F test is very natural. $\|\hat{\mu}\|^2/\|Y - \hat{\mu}\|^2$ measures the size of $\|\hat{\mu}\|^2$ normalized by and estimate of σ^2 so F is unit free. One can also understand $\|\hat{\mu}\|^2$ as quantifying the explained variability in Y and $\|Y - \hat{\mu}\|^2$ as the unexplained variability in Y.

1.3.3 Likelihood ratio statistic

The basic goal of this section is to use our understanding of Gaussian projections to understand why $2 \times$ (log ratio statistic) $= 2\Delta\ell(\hat{\mu})$ is distributed $\chi^2_{\Delta\text{dim}}$ under the null hypothesis where Δdim to denotes the different in dimensions going from the null model to the full model.

To start lets examine the classic linear model

$$Y = \mu + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I).$$

Our hypothesis test is

$$H_0 : \mu \in M_0$$
$$H_1 : \mu \in M_1$$

where $M_0 \subset M_1 \subset V$ are all linear vector spaces. Suppose M_0, M_1 and V have dimensions d_0, d_1 and d respectively. We are free to choose our orthonormal basis of V as

$$\underbrace{\underbrace{\phi_1, \ldots, \phi_{d_0}}_{\text{these span } M_0}, \ \psi_1, \ldots, \psi_{d_1 - d_0}}_{\text{these span } M_1}, \ \xi_1, \ldots \xi_{d_1 - d}.$$

One can use the random coefficients under this basis to generate a log-likelihood for μ given the data as follows

$$\ell(Y|\mu) = -\frac{1}{2\sigma^2}\|Y - \mu\|^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

Now it's clear that the MLE estimates $\hat{\mu}_{M_0}$ and $\hat{\mu}_{M_1}$ just equal the projections $P_{M_0}Y$ and $P_{M_1}Y$ since projections are characterized by the minimizer

of the square norm. Therefore

$$\begin{aligned}
2\Delta\ell(Y|\hat{\mu}) &= 2\ell(Y|P_{M_1}Y) - 2\ell(Y|P_{M_0}Y) \\
&= -\frac{1}{\sigma^2}\|Y - P_{M_1}Y\|^2 + \frac{1}{\sigma^2}\|Y - P_{M_0}Y\|^2 \\
&= -\frac{1}{\sigma^2}\left[\sum_{k=1}^{d_1-d}\langle Y, \xi_k\rangle^2\right] + \frac{1}{\sigma^2}\left[\sum_{k=1}^{d_1-d}\langle Y, \xi_k\rangle^2 + \sum_{k=1}^{d_1-d_0}\langle Y, \psi_k\rangle^2\right] \\
&= \frac{1}{\sigma^2}\sum_{k=1}^{d_1-d_0}\langle Y, \psi_k\rangle^2 \sim \chi^2_{d_1-d_0}, \qquad \text{by Corollary 2}
\end{aligned}$$

The interesting thing is that the distribution $\chi^2_{d_1-d_1}$ persists as an approximation for $2\Delta\ell(\hat{\mu}|Y)$ under much greater generality. The conditions for this approximation are, loosely, that the the statistical model has the local asymptotic normality property (LAN for short), the signal-to-noise ratio of the observations goes to infinity, and the two models M_0 and M_1 are locally linear with local dimension d_0 and d_1 respectively. If one doesn't have the local linear property of M_0, say, then the asymptotic distribution can be a mixture of chi-square random variables. To see why this is true we start with a small extension of the classic linear model

$$Y = \mu + \sigma Z$$

where now $M_0 \subset M_1$ are only assumed locally linear with local dimension d_0 and d_1 respectively (let suppose the embedded space V is still a linear vector space). We will analyze $2\Delta\ell(Y|\hat{\mu})$ in the case that $\sigma \to 0$ to model the situation where the signal-to-noise of the observations goes to infinity. A good example to keep in mind is the case where $V = \mathbb{R}^5$, $M_1 = \mathbb{R}^3$ and M_0 is the two dimensional sphere embedded in M_1.

To analyze the asymptotic distribution of $2\Delta\ell(Y|\hat{\mu})$ in this case suppose the true μ is in M_0. Notice that we are still free to set down a orthonormal basis of V

$$\phi_1, \ldots, \phi_{d_0},\ \psi_1, \ldots, \psi_{d_1-d_0},\ \xi_1, \ldots \xi_{d_1-d}$$

but we do so in such a way that ϕ_k spans the local linear approximation of M_0 around the true μ, and similarly that ϕ_k, ψ_j spans the local linear approximation of M_1 around the true μ. Now as $\sigma \to 0$, the Y gets closer to μ and the local linear approximation of M_0 and M_1 becomes accurate. In this case, the estimates

$$\hat{\mu}_{M_0} \approx P_{L_\mu M_0}Y \text{ and } \hat{\mu}_{M_1} \approx P_{L_\mu M_1}Y$$

where $L_\mu M_0$ and $L_\mu M_1$ denote the local linear approximations at μ. Therefore

$$\begin{aligned} 2\Delta\ell(Y|\hat{\mu}) &= 2\ell(Y|\hat{\mu}_{M_1}) - 2\ell(Y|\hat{\mu}_{M_0}) \\ &\approx 2\ell(Y|P_{L_\mu M_1}Y) - 2\ell(Y|P_{L_\mu M_0}Y) \\ &= \frac{1}{\sigma^2}\sum_{k=1}^{d_1-d_0}\langle Y, \psi_k\rangle^2 \sim \chi^2_{d_1-d_0}. \end{aligned}$$

It now becomes clear what happens when the local linear approximation doesn't hold. For example suppose $V = \mathbb{R}^5$, $M_1 = \mathbb{R}^3$ and $M_0 = \{\|v\|_p = 1 : v \in V\}$ where $p < 1$. In this case M_0 has a kink at the each axis. Suppose the true μ is at the tip of one of those kinks. Then as $\sigma^2 \to 0$, the Y gets closer to μ and the resulting estimate $\hat{\mu}_{M_0}$ starts to behave like the projection of Y onto a half line. This makes the term $\frac{1}{\sigma^2}\|Y - P_{M_0}Y\|^2$ in $2\Delta\ell(Y|\hat{\mu})$ behave like a mixture of chi-squares depending on if the projection of Y lands on the tip of the half line or to the interior of the half line.

Finally, what happens when one doesn't have the nice classical linear model $Y = \mu + \sigma Z$? In this case we observe data, in the form of a vector of observations Y, and we have access only to a log-likelihood surface $\ell(Y|\mu)$. The local asymptotic normality condition essentially says that the MLE $\hat{\mu}$ (computed under model M_1) has the distribution $\hat{\mu} \sim \mathcal{N}(\mu, F^{-1}/n)$ where F is something like the expected Hessian ℓ with one data point. The idea is that, in this case, the likelihood surface of the whole data set $\ell(Y|\mu)$ is well approximated by the likelihood surface of $\hat{\mu}$ under the model $\hat{\mu} \sim \mathcal{N}(\mu, F^{-1}/n)$. This suggested that we can make a new "data vector" $Y_{new} = F^{1/2}\hat{\mu}$ where $Y_{new} = \theta + Z$ and $Z \sim \mathcal{N}(0, n^{-1}I)$. This is the exact situation we were in earlier, with the signal to noise going to infinity, and the chi-square results still hold.

1.4 Gaussian $E(X_0|X_1, \ldots, X_n)$s projection

Most of what we have done with our abstract vector space results could have been don't just as easy (albeit a bit more messy) with coordinates. This hides the true power and generality of the abstract projection results. In this section we analyze projection of random variables which have no natural coordinate vector representation. In particular we show that Gaussian conditional expectation is simply linear projection onto the linear span of the observations.

Suppose $X_0, X_1, \ldots, X_n$ are jointly Gaussian random variables. We can also suppopse, without loss of generality, that $E(X_k) = 0$ for all $k = 0, 1, \ldots, n$. To derive the conditional distribution $X_0|X_1, \ldots, X_n$ notice first

$$E(X_0|X_1, \ldots, X_n) = a_1X_1 + \cdots a_nX_n, \text{ for some constants } a_k.$$

In particular if we let M denote the linear vector space $\{a_1X_1 + \cdots a_nX_n : a_k \in \mathbb{R}\}$ then $E(X_0|X_1, \ldots, X_n) \in M$. The second fact is that $E(X_0|X_1, \ldots, X_n)$ minimizes the mean square distance to X_0.

$$E(X_0|X_1, \ldots, X_n) = \arg\min_{X \in M} E(X_0 - X)^2. \tag{1.4}$$

This establishes that $E(X_0|X_1, \ldots, X_n)$ is a projection. To be specific, let $V = \{a_0X_0 + \cdots a_nX_n : a_k \in \mathbb{R}\}$ be the full set of linear combinations of $X_0, X_1, \ldots, X_n$. This is obviously a linear space of random variables. To make V an abstract vector space we need an inner product for elements of V. The natural inner product is given by covariance:

$$\text{For any } X, Y \in V \text{ define } \langle X, Y \rangle := \text{cov}(X, Y).$$

Notice that the norm $\|X\|$ gives the standard deviation of X. Using this inner product structure it is clear that equation (1.4) gives

$$\boxed{E(X_0|X_1, \ldots, X_n) = P_M X_0.} \tag{1.5}$$

We will use our understanding of projection to derive $E(X_0|X_1, \ldots, X_n)$ but first we mention that, as a consequence of this projection, we have that

$$\boxed{X_0 - P_M X_0 \perp X_1, \ldots, X_n} \tag{1.6}$$

since $X_k \in M$ where $\perp$ in this case means uncorrelated (i.e. independent since we are working with Gaussians). The above fact has an important consequence which will allow us to derive $\text{var}(X_0|X_1, \ldots, X_n)$

$$\boxed{\text{var}(X_0 - P_M X_0) \overset{(1.6)}{=} \text{var}(X_0 - P_M X_0|X_1, \ldots, X_n) = \text{var}(X_0|X_1, \ldots, X_n)}$$

where the last equality follows since $P_M X_0 = E(X_0|X_1, \ldots, X_n)$ is not random given $X_1, \ldots, X_n$. So $\text{var}(X_0|X_1, \ldots, X_n)$ is simply the marginal variance of the residual $X_0 - P_M X_0$. If one has a fast algorithm for computing $E(X_0|X_1, \ldots, X_n)$ and for simulating $X_0, \ldots, X_n$ then one can easily generate a conditional simulation of $X_0|X_1, \ldots, X_n$. In particular since

$X_0|X_1,\ldots,X_n \sim \mathcal{N}(E(X_0|X_1,\ldots,X_n), \mathrm{var}(X_0|X_1,\ldots,X_n))$ to generate a simulation one simply adds a random Z to $E(X_0|X_1,\ldots,X_n)$ where $Z \sim \mathcal{N}(0, \mathrm{var}(X_0|X_1,\ldots,X_n))$. By the above equation we can generate such a Z by simply generating the prediction residual $X_0^* - P_{M^*}X_0^*$ on a new independent simulation $X_0^*,\ldots,X_n^*$.

Now lets use our understanding of projections to compute (1.5). In paricular we need to find and orthonormal basis $\phi_1,\ldots,\phi_n$ of M and then write $P_M X_0 = \sum_{k=1}^n \langle X_0, \phi_k\rangle \phi_k$. Notice that orthonormality in this case means uncorrelated. To uncorrelate $X_1,\ldots,X_n$ simply multiply by $\Sigma_{nn}^{-1/2}$

$$\begin{pmatrix}\phi_1\\ \vdots \\ \phi_n\end{pmatrix} = \Sigma_{nn}^{-1/2}\begin{pmatrix}X_1\\ \vdots \\ X_n\end{pmatrix}.$$

This gives our orthonormal basis of M, the coefficient of which can be computed as

$$\begin{pmatrix}\langle X_0,\phi_1\rangle\\ \vdots \\ \langle X_0,\phi_n\rangle\end{pmatrix} = E\left[\begin{pmatrix}\phi_1\\ \vdots \\ \phi_n\end{pmatrix} X_0\right] = E\left[\Sigma_{nn}^{-1/2}\begin{pmatrix}X_1\\ \vdots \\ X_n\end{pmatrix} X_0\right] = \Sigma_{nn}^{-1/2}\Sigma_{n0}$$

Now

$$\begin{aligned} E(X_0|X_1,\ldots,X_n) = P_M X_0 &= \sum_{k=1}^n \langle X_0,\phi_k\rangle\phi_k \\ &= \begin{pmatrix}\langle X_0,\phi_1\rangle\\ \vdots \\ \langle X_0,\phi_n\rangle\end{pmatrix}^t \begin{pmatrix}\phi_1\\ \vdots \\ \phi_n\end{pmatrix} \\ &= \Sigma_{0n}\Sigma_{nn}^{-1/2}\begin{pmatrix}\phi_1\\ \vdots \\ \phi_n\end{pmatrix} \\ &= \Sigma_{0n}\Sigma_{nn}^{-1}\begin{pmatrix}X_1\\ \vdots \\ X_n\end{pmatrix}. \end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{var}(X_0|X_1,\ldots,X_n) &= \mathrm{var}(X_0 - P_M X_0) \\
&= \mathrm{var}(X_0) - 2\mathrm{cov}(X_0, P_M X_0) + \mathrm{var}(P_M X_0) \\
&= \Sigma_{00} - 2\Sigma_{0n}\Sigma_{nn}^{-1}\Sigma_{n0} + \Sigma_{0n}\Sigma_{nn}^{-1}\Sigma_{nn}\Sigma_{nn}^{-1}\Sigma_{n0} \\
&= \Sigma_{00} - \Sigma_{0n}\Sigma_{nn}^{-1}\Sigma_{n0}
\end{aligned}$$

Chapter 2

Fubini

Motivation:

In a room of people some pairs shake hands and some don't. No two people shake hands more than once and nobody shakes her own hand. Given a person p, $n(p)$ is the number of hands p shook. If the total number of hand shakes is H, then

$$\sum_p n(p) = 2H$$

Given a person p and hand shake h, define $\mathbb{I}_{ph} = 1$ if person p participated in handshake h, and it is zero otherwise. Then note that $n(p) = \sum_h \mathbb{I}_{ph}$

$$\sum_p n(p) = \sum_p \sum_h \mathbb{I}_{ph} = \sum_h \sum_p \mathbb{I}_{ph} = \sum_h 2 = 2H$$

!!! If 5 people in a room each claim to have shaken 3 hands, then someone is lying. (15 is not an even number)

FUBINI: (Infinite Sums)

If $\sum_{j,k} |a_{jk}| < \infty$, then $\sum_{j,k} a_{jk}$ exists and

$$\sum_{j,k} a_{jk} = \sum_j \sum_k a_{jk} = \sum_k \sum_j a_{jk}$$

Consider $a_{jk} = a^+_{jk} - a^-_{jk}$, Then Fubini theorem holds if one of the below holds

- $\sum_j \sum_k a_{jk}^+ < \infty$
- $\sum_j \sum_k a_{jk}^- < \infty$
- $\sum_k \sum_j a_{jk}^+ < \infty$
- $\sum_k \sum_j a_{jk}^- < \infty$

Example

Consider a_{ij} as follows

i/j	1	2	3	4	5
1	1	-1	0	0	0
2	0	1	-1	0	0
3	0	0	1	-1	0
4	0	0	0	1	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$a_{jk} = \begin{cases} 1 & , \quad j = k \\ -1 & , \quad j = k+1 \\ 0 & , \quad \text{o'wise} \end{cases}$$

Note that $\sum_j a_{jk} = 0$, hence $\sum_k \left(\sum_j a_{jk} \right) = 0$. However, if $j \geq 2$ $\sum_k a_{jk} = 0$ and when $j = 1$ $\sum_k a_{jk} = 1$, hence $\sum_j \left(\sum_k a_{jk} \right) = 1$. Thus, here Fubini FAILS

$$0 = \sum_k \sum_j a_{jk} \neq \sum_j \sum_k a_{jk} = 1$$

Here in this example $\sum_k \sum_j a_{jk}^+ = \infty$ and $\sum_k \sum_j a_{jk}^- = \infty$, thus $\sum_k \sum_j a_{jk} = \sum_k \sum_j a_{jk}^+ - \sum_k \sum_j a_{jk}^- = \infty - \infty$ (Undefined). That is the main condition to swap the sums do not hold, i.e., $\sum_k \sum_j |a_{jk}| < \infty$

Example

Let $A_1, A_2, ...$ be sequence of events, define $N = \sum_{i=1}^{\infty} \mathbb{I}_{A_i}$.

$$\mathbb{E}N = \mathbb{E}\left(\sum_{i=1}^{\infty} \mathbb{I}_{A_i} \right) = \sum_{i=1}^{\infty} \mathbb{E}\mathbb{I}_{A_i} = \sum_{i=1}^{\infty} P(A_i)$$

The reason we were able to swap the Expectation and the sum is that it satisfies the Fubini criteria, i.e.,

$$\sum_{i=1}^{\infty} P(A_i) < \infty \rightarrow \quad \mathbb{E}N < \infty \rightarrow \quad P(N = \infty) = 0 \rightarrow \quad P(A_i's \text{ happening i.o}) = 0$$

Which is the first Cantelli Lemma.

Original Fubini Theorem:

Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be two complete σ-finite measure spaces. Suppose f is integrable on $(X \times Y)$, then

$$\begin{aligned}
&(i) \int_Y f(x,y)d\nu(y) \text{ is integrable on X} \\
&(ii) \int_X f(x,y)d\mu(y) \text{ is integrable on Y} \\
&(iii) \int_X \int_Y f(x,y)d\nu d\mu = \int_{X\times Y} f(x,y)d(\mu \otimes \nu) = \int_Y \int_X f(x,y)d\mu d\nu
\end{aligned} \tag{2.1}$$

To apply Fubini, it requires that f is integrable on $X \times Y$

$$\begin{aligned}
&(i) \int_Y \left(\int_X |f|d\mu \right) d\nu(y) < \infty \\
&(ii) \int_X \left(\int_Y |f|d\nu \right) d\mu(y) < \infty \\
&(iii) \int_{X\times Y} |f|d(\mu \otimes \nu) < \infty
\end{aligned} \tag{2.2}$$

Chapter 3

Ideas Involving Bayesian Risk

In this chapter we will explore some ideas involved in finding a Bayes rule (or Bayes estimator), and its associated Bayes risk under a specific loss function.

Consider the family of Poisson distributions $\mathcal{P}(\lambda)$, $\lambda > 0$, with p.m.f. given by

$$p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots,$$

where we are interested in estimating λ under the loss function

$$l(\lambda, \delta(X)) = \frac{(\lambda - \delta(X))^2}{\lambda}. \tag{3.1}$$

Claim 5. *The Gamma family of distributions forms a conjugate family of priors for λ.*

Specifically, let $\lambda \sim Gamma(\alpha, \beta)$ be a prior on λ with prior hyper parameters $\theta_0 = (\alpha, \beta)$. Then the posterior is $\lambda|x \sim Gamma\left(\alpha + x, \frac{\beta}{\beta+1}\right)$, with posterior hyper parameters $\theta_1 = \left(\alpha + x, \frac{\beta}{\beta+1}\right)$.

Under this Gamma prior, a natural question is to find the associated Bayes estimator, or Bayes rule, $\widehat{\lambda}_\pi$.

By definition, a Bayes estimator $\delta^*(X)$ is a decision rule such that it minimizes the posterior risk; i.e.

$$r(\delta^*(x)|x) = \inf_\delta \ r_\pi(\delta|x)$$

where $r_\pi(\delta|x)$ is defined as

$$r_\pi(\delta|x) = \mathbb{E}_{\lambda|x}(l(\lambda, \delta(X))|X = x) = \int_\Lambda l(\lambda, \delta(X))\pi(\lambda|x)\, d\lambda, \tag{3.2}$$

and the expectation is taken with respect to the posterior distribution of λ given $X = x$, and λ is distributed as π.

In order to find this desired rule, $\widehat{\lambda}_\pi = \delta^*(X)$, and subsequently its associated Bayes risk, there are a number of ways to go about performing the actual computations. Our goal in this note is to show that not all of these methods are "created equal" in terms of computational time, which is especially relevant when taking an in-class exam.

Method I: Finding $\widehat{\lambda}_\pi$ directly.
Using equation (3.2), we can substitute in our loss function (3.1), and arrive at

$$r_\pi(\delta|x) = \mathbb{E}_{\lambda|x}\left(\frac{(\lambda-\delta)^2}{\lambda}\right) = \mathbb{E}_{\lambda|x}(\lambda) - 2\delta + \delta^2\mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right)$$

We can find the infimum of this Bayes risk over all decision rules δ by taking the derivative with respect to δ, and setting equal to 0.

$$\frac{\partial r_\pi(\delta|x)}{\partial\delta} = -2\delta + \delta^2\mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right)$$

$$\implies \quad \delta^*(X) = \frac{1}{\mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right)}$$

This expected value, taken under the posterior distribution, is equal to $\frac{\beta+1}{(\alpha+x-1)\beta}$

$$\begin{aligned}
\mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right) &= \int \frac{1}{\lambda}\pi(\lambda|x)\,d\lambda = \int \frac{1}{\lambda}\frac{1}{\Gamma(\alpha+x)\left(\frac{\beta}{\beta+1}\right)^{\alpha+x}}\lambda^{\alpha+x-1}e^{-(\frac{\beta+1}{\beta})\lambda}\,d\lambda \\
&= \frac{\Gamma(\alpha+x-1)\left(\frac{\beta}{\beta+1}\right)^{\alpha+x-1}}{\Gamma(\alpha+x)\left(\frac{\beta}{\beta+1}\right)^{\alpha+x}}\int \frac{\lambda^{(\alpha+x-1)-1}e^{-(\frac{\beta+1}{\beta})\lambda}}{\Gamma(\alpha+x-1)\left(\frac{\beta}{\beta+1}\right)^{\alpha+x-1}}\,d\lambda \\
&= \frac{1}{(\alpha+x-1)\left(\frac{\beta}{\beta+1}\right)}
\end{aligned}$$

Thus,

$$\delta^*(X) = \widehat{\lambda}_\pi = \frac{(\alpha+x-1)\beta}{\beta+1}$$

Method II: A second method for finding this Bayes rule utilizes the following well-known theorem, and a subsequent corollary.

Theorem 1. *Given a squared-error loss function $l(\lambda, \delta(X)) = (\lambda - \delta(X))^2$, and a prior π on λ, the Bayes rule is found by taking the expected value of the associated posterior distribution. i.e. $\widehat{\lambda}_\pi = E_{\lambda|x}(\lambda)$.*

Corollary 3. *Let $\omega > 0$ be a positive 'weight'. Then, for the so-called 'weighted squared-error' loss function $l(\lambda, \delta(X)) = \frac{(\lambda - \delta(X))^2}{\omega}$. The Bayes rule is found by solving the following integral:*

$$\widehat{\lambda}_\pi = \int_\Lambda \frac{(\lambda - \delta(X))^2}{\omega} \pi(\lambda|x) d\lambda$$

where π is a prior on λ, and $\pi(\lambda|X)$ is the density of the associated posterior distribution.

For our loss function given in (3.1) we recognize that we have a "weighted" square-error loss. Thus, we identify $\omega = \lambda$, and write

$$\widehat{\lambda}_\pi = \int_\Lambda \frac{(\lambda - \delta(X))^2}{\lambda} \pi(\lambda|x) d\lambda$$

Now, since our posterior distribution is a gamma with parameters $\left(\alpha + x, \frac{\beta}{\beta+1}\right)$, then taking this posterior density and dividing by λ gives a density of a gamma distribution with parameters $\left(\alpha + x - 1, \frac{\beta}{\beta+1}\right)$. It follows that the Bayes rule is simply the expected value of this newly-formed gamma distribution. i.e.

$$\widehat{\lambda}_\pi = \frac{(\alpha + x - 1)\beta}{\beta + 1}.$$

Next, we are interested in finding the Bayes risk of this Bayes rule. i.e. $r(\pi, \widehat{\lambda}_\pi)$.

The Bayes risk for any decision rule δ is defined as

$$r(\pi, \delta) = \mathbb{E}_\lambda(R(\lambda, \delta)) = \mathbb{E}_\lambda \left[\mathbb{E}_{x|\lambda}\theta \left(l(\lambda, \delta)\right)\right]$$

or, by explicitly writing out the expectations,

$$r(\pi, \delta) = \int R(\theta, \delta)\pi(\lambda)\, d\lambda = \int \left[\int l(\lambda, \delta) p(x|\lambda)\, dx\right] \pi(\lambda)\, d\lambda$$

Note that with $\widehat{\lambda}_\pi = (\alpha + x - 1)\frac{\beta}{\beta+1}$, we can write

$$\widehat{\lambda}_\pi = \gamma x + (\alpha - 1)\gamma, \quad \text{where} \quad \gamma = \frac{\beta}{\beta + 1}.$$

Method I: Compute directly

$$\begin{aligned} r\left(\pi,\widehat{\lambda}\right) &= \mathbb{E}_\lambda\left[\mathbb{E}_{x|\lambda}\left(l(\lambda,\widehat{\lambda})\right)\right] = \mathbb{E}_\lambda\left[\mathbb{E}_{x|\lambda}\left(\frac{(\lambda-\widehat{\lambda})^2}{\lambda}\right)\right] \\ &= \mathbb{E}_\lambda\left[\mathbb{E}_{x|\lambda}\left(\lambda - 2\widehat{\lambda} + \frac{1}{\lambda}\widehat{\lambda}^2\right)\right] = \mathbb{E}_\lambda\left[\lambda - 2\mathbb{E}_{x|\lambda}\left(\widehat{\lambda}\right) + \frac{1}{\lambda}\mathbb{E}_{x|\lambda}\left(\widehat{\lambda}^2\right)\right] \\ &= \cdots \end{aligned}$$

This method quickly becomes a computational nightmare.

Method II: Compute using the MSE: $\mathbb{E}(\lambda-\widehat{\lambda})^2 = Var(\widehat{\lambda}) + [E(\widehat{\lambda})-\lambda]^2$

$$\begin{aligned} r\left(\pi,\widehat{\lambda}\right) &= \mathbb{E}_\lambda\left[\mathbb{E}_{x|\lambda}\left(l(\lambda,\widehat{\lambda})\right)\right] = \mathbb{E}_\lambda\left[\frac{1}{\lambda}\left(Var_{x|\lambda}(\widehat{\lambda}) + [\mathbb{E}_{x|\lambda}(\widehat{\lambda})-\lambda]^2\right)\right] \\ &= \mathbb{E}_\lambda\left[\frac{1}{\lambda}\left(\gamma^2\lambda + [\gamma\lambda + (\alpha-1)\gamma - \lambda]^2\right)\right] \\ &= \mathbb{E}_\lambda\left[\frac{1}{\lambda}\left(\gamma^2\lambda + \lambda^2(\gamma-1)^2 + 2\lambda\gamma(\gamma-1)(\alpha-1) + \gamma^2(\alpha-1)^2\right)\right] \\ &= \mathbb{E}_\lambda\left[\gamma^2 + \lambda\frac{\gamma^2}{\beta^2} - 2\frac{\gamma^2}{\beta}(\alpha-1) + \frac{1}{\lambda}\gamma^2(\alpha-1)^2\right] \\ &= \gamma^2 + \alpha\beta\frac{\gamma^2}{\beta^2} - 2\frac{\gamma^2}{\beta}(\alpha-1) + \frac{1}{(\alpha-1)\beta}\gamma^2(\alpha-1)^2 \\ &= \gamma^2 + \alpha\frac{\gamma^2}{\beta} - \frac{\gamma^2}{\beta}(\alpha-1) = \gamma^2 + \frac{\gamma^2}{\beta} = \gamma^2\left(1+\frac{1}{\beta}\right) = \frac{\beta}{\beta+1} \qquad \square \end{aligned}$$

where we used the following:

$$Var_{x|\lambda}(\widehat{\lambda}) = \gamma^2\lambda, \quad \mathbb{E}_{x|\lambda}(\widehat{\lambda}) = \gamma\lambda + (\alpha-1)\gamma, \quad \gamma - 1 = \frac{-\gamma}{\beta}$$

Method III: Compute by swapping expectations.
Switch the order of expectations, which is allowed by applying Bayes Theorem to change from the prior to the posterior, and Fubini's Theorem which allows us to swap the order of integration.

$$r(\pi,\delta) = \int\left[\int l(\lambda,\delta)p(x|\lambda)\,dx\right]\pi(\lambda)\,d\lambda = \int\left[\int l(\lambda,\delta)\pi(\lambda|x)\,d\lambda\right]p(x)\,dx = \mathbb{E}_x\left[\mathbb{E}_{\lambda|x}\left(l(\lambda,\widehat{\lambda})\right)\right]$$

Then,

$$\begin{aligned} r\left(\pi, \widehat{\lambda}\right) &= \mathbb{E}_x\left[\mathbb{E}_{\lambda|x}\left(l(\lambda, \widehat{\lambda})\right)\right] = \mathbb{E}_x\left[\mathbb{E}_{\lambda|x}\left(\lambda - 2\widehat{\lambda} + \frac{1}{\lambda}\widehat{\lambda}^2\right)\right] \\ &= \mathbb{E}_x\left[\widehat{\lambda} + \frac{\beta}{\beta+1} - 2\widehat{\lambda} + \frac{1}{\widehat{\lambda}}\widehat{\lambda}^2\right] = \mathbb{E}_x\left[\frac{\beta}{\beta+1}\right] \\ &= \frac{\beta}{\beta+1} \quad \square \end{aligned}$$

where we used

$$\mathbb{E}_{\lambda|x}(\lambda) = \frac{(\alpha+x)\beta}{\beta+1} = \frac{(\alpha+x-1)\beta}{\beta+1} + \frac{\beta}{\beta+1} = \widehat{\lambda} + \frac{\beta}{\beta+1}, \text{ and } \mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right) = \frac{1}{\widehat{\lambda}}.$$