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*

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

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 $d_1, \dots, d_d$ 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 $\phi_1, \dots, \phi_d$ is a orthonormal basis of V then*

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

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

Corollary 2. *Suppose $\phi_1, \dots, \phi_d$ 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 $\phi_1, \dots, \phi_m$. If $Y = \sum_{k=1}^d c_k \phi_k$ 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) $var(\langle v, Y\rangle) = var(v_1 c_1 + \cdots v_d c_d) = \sigma^2 \|v\|^2$.

Chapter 2

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

$$l(\lambda, \delta(X)) = \frac{(\lambda - \delta(X))^2}{\lambda}. \tag{2.1}$$

Claim 4. *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

$$r_\pi(\delta|x) = \mathbb{E}_{\lambda|x}(l(\lambda, \delta(X))|X = x) = \int_\Lambda l(\lambda, \delta(X))\pi(\lambda|x)\, d\lambda, \tag{2.2}$$

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 (2.2), we can substitute in our loss function (2.1), 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.

$$\begin{aligned} \frac{\partial r_\pi(\delta|x)}{\partial \delta} &= -2\delta + \delta^2 \mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right) \\ \implies \quad \delta^*(X) &= \frac{1}{\mathbb{E}_{\lambda|x}\left(\frac{1}{\lambda}\right)} \end{aligned}$$

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 (2.1) 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

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

Chapter 3

Linear Regression and Gauss Markov Theorem in terms of Orthogonal Projection

In this chapter, we use the material in chapter 1 to prove the Gauss-Markov Theorem, and we interpret the Least Squares Estimator from the perspective of Orthogonal Projection.

3.1 The Gauss Markov Theorem

$$y_{n\times 1} = \mu_{n\times 1} + \epsilon_{n\times 1}$$

where $\mu \in \mathcal{M} \subset \mathcal{R}^n,\ and\ \epsilon \sim N(0, \sigma^2 I_{n\times n})$

The Gauss Markov Theorem says, for any $v \in \mathcal{R}^n$, $< v, P_M y >$ is the BLUE of $< v, \mu >$.

3.2 Proof

- $< v, P_M y >$ is linear in y. Let $P_M^{\perp}$ be the Projection to the subspace $\mathcal{M}^{\perp}$ of $\mathcal{R}^n$ that is orthogonal to $\mathcal{M}$. Then for any $v \in \mathcal{R}^n$ we can decompose v in to $v = P_M v + P_M^{\perp} v$

  $$< v, P_M Y >=< P_M v + P_M^{\perp} v, P_M Y >=< P_M v, P_M Y > + < P_M^{\perp} v, P_M Y >=< P_M v, Y >$$

- $< v, P_M Y >$ is unbiased.

$$E < v, P_M Y > = E < P_M v, Y > = < P_M v, \mu > = < v, \mu >, \forall \mu \in \mathcal{M}$$

- $< v, P_M y >$ has minimum variance among all linear unbiased estimators of $< v, \mu >$

  Let $< x, y >$ be a different linear unbiased estimator of $< v, P_M y >$.

  $\rightarrow E < x, y > = < x, Ey > = < x, \mu > = < v, \mu > \; \forall \mu \in \mathcal{M}$
  $\rightarrow < P_M x, \mu > = < P_M v, \mu > \; \forall \mu \in \mathcal{M}$
  $\rightarrow P_M x = P_M v$

$$\begin{aligned} Var < x, y > &= Var < x, \epsilon > \\ &= Var < \sum < x, \phi_k > \phi_k, \epsilon > \\ &= Var(\sum < x, \phi_k >< \phi_k, \epsilon >) \\ &= \sum < x, \phi_k >^2 \sigma^2 \\ &= \sum_{k=1}^{dim(\mathcal{M})} < x, \phi_k >^2 \sigma^2 + \sum_{k=dim(\mathcal{M})+1}^{n} < x, \phi_k >^2 \sigma^2 \\ &= ||P_M x||^2 \sigma^2 + ||P_M^{\perp} x||^2 \sigma^2 \\ &= ||P_M v||^2 \sigma^2 + ||P_M^{\perp} x||^2 \sigma^2 \end{aligned}$$

$$\begin{aligned} Var < v, P_M y > &= Var < P_M v, \epsilon > \\ &= Var < \sum < P_M v, \phi_k > \phi_k, \epsilon > \\ &= Var(\sum < P_M v, \phi_k >< \phi_k, \epsilon >) \\ &= \sum < P_M v, \phi_k >^2 \sigma^2 \\ &= \sum_{k=1}^{dim(\mathcal{M})} < P_M v, \phi_k >^2 \sigma^2 + \sum_{k=dim(\mathcal{M})+1}^{n} < P_M v, \phi_k >^2 \sigma^2 \\ &= ||P_M v||^2 \sigma^2 + 0 \end{aligned}$$

$$Var < x, y > - Var < v, P_M y > = ||P_M^{\perp}||^2 \sigma^2 \geq 0$$

3.3 LSE

$$Y = X\beta + \epsilon$$

Let $\mu = X\beta$, then $\mu \in Col(X)$. BLUE of $\mu\ is\ \hat{\mu} = X(X^tX)^{-1}X^tY,\ and\ \hat{\mu}$ is a orthogonal projection of Y onto $\mathcal{M} = Col(X)$. Orthogonalize the columns of X so that

$$X = [\phi_1|\ldots|\phi_p]$$

then, $\phi_1, \ldots, \phi_p$ is an orthonormal basis of $\mathcal{M}$. Then one can write

$$\begin{aligned}
\hat{\mu} = P_My = XX^tY &= [\phi_1|\ldots|\phi_n] \begin{bmatrix} \phi_1^t \\ \vdots \\ \phi_p^t \end{bmatrix} Y \\
&= [\phi_1|\ldots|\phi_n] \begin{bmatrix} <\phi_1, Y> \\ \vdots \\ <\phi_p, Y> \end{bmatrix} \\
&= \sum_{k=1}^{p} <\phi_k, Y> \phi_k
\end{aligned}$$

3.4 Testing

$H_0 : \beta = 0$

Let $T = ||Y - \hat{\mu}||^2$ be the test statistic.

$$\begin{aligned}
T &= ||Y - P_MY||^2 \\
&= ||P_M^{\perp}Y||^2 \\
&= \sum_{k=p+1}^{n} <Y, \phi_k>^2
\end{aligned}$$

Under H_0, $Y \sim N(0, \sigma^2 I_n)$, thus $\frac{T}{\sigma^2} \sim \chi^2_{n-p}$

3.5 Residual Plots

Residuals $= Y - \hat{\mu} = Y - P_M Y = P_M^{\perp} Y \perp P_M Y$