MATRIX ALGEBRA IN STATISTICS: ADDITIONAL TOPICS

BASED ON LECTURES BY PROF.S.MACEACHERN AND **NOTES BY H.LUO**

"Enjoy it." thus said Prof.S.Maceachern.

Date: 2016-03-10.

Contents

[1][Ghosh et.al]

1. Matrix Operations

1.1. **Miscellaneous.** One thing to remember is that the matrix operation is usually not directly related to the sample space yet the image of a random variable space.

Definition 1. (Commutative matrices) Two matrices $\mathbf{A}, \mathbf{B}$ are called commutative matrices if $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$. We know that identity matrices and diagonal matrices commute with any other matrices. The question is what is a general condition that two matrices commute? A matrix commute with its inverse. But a sufficient and necessary condition for $\{\mathbf{A}_1, \cdots \mathbf{A}_n\}$ mutually commute with each other is that there exists a matrix $\mathbf{U}$ such that $\mathbf{UA_iU}$ are all of diagonal forms. (which is also known as *simultaneous diagonalization*.)

Definition 2. (Row/Column sums) We can write the row sum or column sum as matrix multiplication, like $\mathbf{AJ}$ is the row sum while the $\mathbf{JA}$ is the column sum. The vector $\mathbf{J}$ is the all-one vector of corresponding dimension.

Definition 3. (Inverse product rule and transpose product rule) $(\mathbf{AB})^{-1} = \mathbf{B^{-1}A^{-1}}$,$(\mathbf{AB})' = \mathbf{B'A'}$,$\left(\mathbf{A}'\right)^{-1} = \left(\mathbf{A^{-1}}\right)'$

Definition 4. (Triangular matrices) Upper/lower triangular matrices are closed under addition and matrix multiplication, yet an operation between an upper and a lower triangular matrix has nothing special.

Definition 5. (Pre-/Post-multiplication) Premultiply $\mathbf{A}$ by a matrix $\mathbf{P}$ is equivalent to manipulate the rows of this matrix. Postmultiply $\mathbf{A}$ by a matrix $\mathbf{P}$ is equivalent to manipulate the columns of this matrix. If the $\mathbf{P}$ happens to be a permutation matrix, then this pre-/post-multiplication can be used in cross-validation, permutation test and many places in statistics.

An example of this is the *permutation matrix* whose columns are consisting of the columns of identity matrix. Premultiplying permutation matrix permutes the rows of the matrix; Postmultiplying permutation matrix permutes the columns of the matrix. Permutation matrices can be used in resampling, cross-validation and sometimes randomization. For example, See [Weisberg] Chap.7 for case resampling bootstrap method.

Theorem 6. *Any permutation matrix is a composition of* reflection *and* rotation *matrix.*

1.2. **Markov Chain.**

Definition 7. (Markov chain) If a random variable sequence $\mathbf{X}_i$ satisfies following conditions, then we call it a *Markov chain*[2]:

(i) (finite states)$\mathbf{X}_i \in \{1, \cdots, s\}, \forall i$

(ii) (Markov property, of length 1) $\mathbf{X}_n|\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \cdots \mathbf{X}_1 \equiv \mathbf{X}_n|\mathbf{X}_{n-1}$ this conditional distribution is called a *transitional distribution* from $\mathbf{X}_{n-1}$ to $\mathbf{X}_n$.

We can write the probability p_{ij} from $\mathbf{X}_{n-1} = i$ to $\mathbf{X}_n = j$ into the matrix form. This matrix is called a *transitional matrix* $\mathbf{P}_{\mathbf{X_n},\mathbf{X_{n-1}}} = \{p_{ij}\}$. It is not hard to see that the $i-th$ row vector of a transitional matrix $\mathbf{P}_{\mathbf{X_n},\mathbf{X_{n-1}}}$ is the conditional distribution $\mathbf{P}_{\mathbf{X_n},\mathbf{X_{n-1}=i}}$. And if all the transitional matrix are the same and independent of the (time) parameter n, we usually denote it using the same notation $\mathbf{P}$.

The entries of this transition matrix satisfies $\sum_i p_{ij} = 1, p_{ij} \geq 0$.

Given an *initial state* $\pi_\mathbf{0}$, the state at time t can be calculated using matrix multiplication as $\pi_\mathbf{0}'\mathbf{P}^t$. The *limiting distribution* is defined to be $\mathbf{P}_\infty = lim_{t\to\infty}\mathbf{P}^t$. This Markov model is over-simplified, to improve its accuracy, we might want to increase the dependence length or include more kinds of states. One method is to increase the length of the state-chain.

For the 1-step dependence model, the limiting distribution can have basically three types of limiting distributions:

(i) Periodic behavior: $\pi_1 \to \pi_2 \to \pi_3 \to \pi_1 \to \cdots$.

(ii) Absorbing behavior: $\pi_1 \to \pi_2 \to \pi_3 \to \pi_3 \to \cdots$.

(iii) Limiting behavior: $\pi_1 \to \pi_2 \to \pi_3 \to \cdots \to \pi_\infty$.

The idea of *Markov Chain Monte-Carlo* method is to build a Markov chain with a specific limiting distribution and to simulate samples from this Markov chain.

1.3. **ANOVA.** $\mathbf{X} = \begin{pmatrix} 1 & 1 & \\ 1 & 1 & \\ 1 & & 1 \\ 1 & & 1 \\ 1 & & 1 \end{pmatrix}$ is a design matrix of a one-way ANOVA model $Y = \alpha_i + \epsilon, i = 1, 2$ with only 1 main-effect with 2 levels. $\mathbf{X} = \begin{pmatrix} 1 & 1 & & 1 & & \\ 1 & 1 & & & 1 & \\ 1 & 1 & & & & 1 \\ 1 & & 1 & 1 & & \\ 1 & & 1 & & 1 & \\ 1 & & 1 & & & 1 \end{pmatrix}$ is a design matrix of a two-way ANOVA model $Y = \alpha_i + \beta_j + \epsilon, i = 1, 2, j = 1, 2, 3$

[2]Markov chain is widely used in forecasting, natural language processing and other situations.

with 2 main-effect with 2/3 levels. $\mathbf{X} = \begin{pmatrix} 1 & 1 & & 1 & & 1 & & & \\ 1 & 1 & & & 1 & & 1 & & \\ 1 & 1 & & & & 1 & & 1 & \\ 1 & & 1 & 1 & & & & & 1 \\ 1 & & 1 & & 1 & & & & \\ 1 & & 1 & & & 1 & & & \end{pmatrix}$ is a design matrix of a two-way ANOVA COMPLETE model $Y = \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon, i = 1, 2, j = 1, 2, 3$ with 2 main-effect with 2/3 levels. All of these models only replicates once, if you want more, you can try increase the number of rows. Similarly, we can also write ANCOVA model into a design matrix. However, we must impose some assumptions on the interaction term since if we do not, then the continuous variable will require infinitely many columns to represent, which is not possible.

The advantage is that if we want to decide whether a contrast is estimable, then it suffices to write this contrast using the basis from $\mathcal{C}(\mathbf{X})$. And the identifiability relies on whether the column vectors of the design matrix form a basis of $\mathcal{C}(\mathbf{X})$.

1.4. **Reparameterization.** $\mathbf{Y} = \mathbf{X}\beta + \epsilon = [\mathbf{XA}]\,[\mathbf{A^{-1}}\beta] + \epsilon$. $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{XAA^{-1}}) \subset \mathcal{C}(\mathbf{XA}) \Rightarrow \mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{XA})$ so such a reparameterization will not affect the model. We should also notice that the *identifiability* of the parameter β in such a model is equivalent to the full column rank condition on the design matrix $\mathbf{X}$. In ANOVA setting, the parameter is generally not identifiable.

Therefore reparameterization is equivalent to change of basis in the $\mathcal{C}(\mathbf{X})$. In statistics, we might want to proceed a change of basis for many purposes:

(i)Better understanding of the model. For example, sometimes we might want to take a control versus treatment comparison.

For example in a single-replicate two sample example, $\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ may be called an *effective coding*, while the $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ may be a *control-versus-treatment* setting.

(ii)Mathematical convenience. For example, the Gram-Schmidt orthogonalization procedure.

(iii)Computational speed for design of algorithms.

1.5. **Partitioned matrices.** We know two matrices with the same blocking structures can be added, scalaring and multiplying block-wisely. There is a scene where the block matrix comes into play in statistics. Consider a random vector's covariance matrix, if the covariance matrix is of the block-diagonal form, that might indicate the uncorrelated of two groups of components of this random vector. Moreover, if the random vector is distributed as a mixed model setup, then the components in the random vector corresponding to the same block will share the same random effect.

There are two ways of regarding a model as a multivariate distribution, one is to fix one component of the random vector as response and treat others as regressors, this is the popular view. In this view, we regard fitting the model as finding the conditional distribution of $\mathbf{y}|\mathbf{x}$. Another view is to view all components in the random vectors symmetrically. This view helps us to understand the correlation coefficients' distribution better. [3].

Partitioned matrices can also be used to describe the situation where we drop a regressor from our linear model. Each time we drop a group of regressors from our model we take off the corresponding columns in our design matrix in the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.

For the proof of results on partitioned matrices, there are generally two kinds of argument, one is to decompose a partitioned matrix into a product of a few simpler partitioned matrices with special properties like block-diagonal; the other is to discuss the basis consisting of columns of the partitioned blocks.

[3][Anderson] Chap.3

2. The Rank and Norm

2.1. Rank and dimension.

Definition 8. (Dimension) The dimension of a linear space is the cardinality of the minimal spanning set/maximal linearly independent set. And two linear spaces are equivalent iff they are of the same dimension.

Definition 9. (Rank) The rank of a matrix is the dimension of its column/row space. There are a few rank inequalities which are of some use:

Theorem 10. *Followings are useful inequalities for rank of matrices:*

Partitioned Inequality. $max(rank(\mathbf{A}), rank(\mathbf{B})) \leq rank(\mathbf{A}\vdots\mathbf{B}) \leq rank(\mathbf{A}) + rank(\mathbf{B})$ *The equality holds iff the columns of two matrices are linearly independent.*

Sylvester Inequality. $rank(\mathbf{AB}) \leq rank(\mathbf{A}) + rank(\mathbf{B}) \leq rank(\mathbf{AB}) + n$ *The equality holds iff the two matrices are of full rank.*

Frobenius Inequality. $rank(\mathbf{AB}) + rank(\mathbf{BC}) \leq rank(\mathbf{ABC}) + rank(\mathbf{B})$ *The equality holds iff the three matrices are of full rank.*

Moreover, if we treat the matrix as a linear mapping defined by $x \mapsto \mathbf{A}x$, then the rank of the matrix is exactly the dimension of the range of this linear mapping.

Trace. [4]The trace of matrices, being the sum of all diagonal terms, is invariant under *cyclic permutations*. $trace(\mathbf{ABC}) = trace(\mathbf{BCA}) = trace(\mathbf{CAB})$.

Theorem 11. *For an* ***idempotent matrix*** $\mathbf{A}$, *we have identity* $rank(\mathbf{A}) = trace(\mathbf{A})$. *This can be shown by using the rank factorization and the sufficient-necessary condition for an existence of right/left inverse. OR we can think that the trace is the sum of eigenvalues.*

2.2. Matrix norms.

Definition 12. (Norms) [5]A norm should satisfy

(1) (positivity)$\|\mathbf{A}\| \geq 0$, and the equality holds iff$\mathbf{A} = \mathbf{0}$.
(2) (scalar) $\|\alpha\mathbf{A}\| = |\alpha| \|\mathbf{A}\|$.
(3) (triangle inequality) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

Definition 13. (Distances) A distance should satisfy

(1) (positivity) $d(\mathbf{A}, \mathbf{B}) \geq 0$, and the equality holds iff $\mathbf{A} = \mathbf{B}$.

[4]Here we briefly mentioned two information criteria, which is AIC and BIC. This is because we knew that for an idempotent matrix, $trace(\mathbf{A}) = rank(\mathbf{A})$, and hat matrix in the linear model regression setting happens to be idempotent. So the information criteria can naturally be derived as a measurement, other than ranks, of the size of the model. See [Ghosh et.al] Chap.5 and [Ando].

[5]The norm of the hat matrix can be used to measure how far the fitted model deviates from the "true" model, we can discuss consistency and multicolinearity using the matrix norm concept.

(2) (reflexivity) $d(\mathbf{A},\mathbf{B}) = d(\mathbf{B},\mathbf{A})$.
(3) (triangle inequality) $d(\mathbf{A},\mathbf{B}) \leq d(\mathbf{A},\mathbf{C}) + d(\mathbf{C},\mathbf{B})$.

$\mathbf{L}^1$ norm. The maximum absolute column sum of the matrix.
$\mathbf{L}^2$ norm. $\sqrt{\lambda_{max}(\mathbf{A}'\mathbf{A})}$
$\mathbf{L}^{p,q}$ norm. $\frac{\|\mathbf{Ax}\|_p}{\|\mathbf{x}\|_q}$
$\mathbf{L}^\infty$ norm. The maximum absolute row sum of the matrix.
Frobenius norm. $\sqrt{trace(\mathbf{A}'\mathbf{A})}$

Theorem 14. ***Cauchy-Schwartz Inequality for norms.*** $\|\mathbf{x}\cdot\mathbf{y}\|_{\mathbb{R}^1} \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$ *the equality holds iff* $\mathbf{x}\cdot\mathbf{y} = 0$.

2.3. **Inner products.**

Definition 15. (Inner products) An inner product on a vector space is a scalar function $\langle\cdot,\cdot\rangle : V \times V \mapsto \mathbb{C}$ such that:

(1) (conjugacy) $\langle\mathbf{x},\mathbf{y}\rangle = \overline{\langle\mathbf{y},\mathbf{x}\rangle}$
(2) (linearity in the first component) $\langle a\mathbf{x_1} + b\mathbf{x_2},\mathbf{y}\rangle = a\langle\mathbf{x_1},\mathbf{y}\rangle + b\langle\mathbf{x_2},\mathbf{y}\rangle, \forall a,b \in \mathbb{C}$
(3) (positivity) $\langle\mathbf{x},\mathbf{x}\rangle \geq 0$, and the equality holds iff $\mathbf{x} = 0$.

Generally speaking, when we have a symmetric positive definite matrix $\mathbf{W}$, we can firstly define an inner product by $\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{W}} \coloneqq \mathbf{x}'\mathbf{W}\mathbf{y}$, then as usual the inner product will induce corresponding norm and distance such that $\|\mathbf{x}\|_{\mathbf{W}} \coloneqq \sqrt{\langle\mathbf{x},\mathbf{x}\rangle_{\mathbf{W}}} = \sqrt{\mathbf{x}'\mathbf{W}\mathbf{x}}, d(\mathbf{x},\mathbf{y}) \coloneqq \|\mathbf{x}-\mathbf{y}\|_{\mathbf{W}}$. Using the Cholesky decomposition of positive definite matrix, we can also regarded such an inner product defined above as the usual inner product of two transformed vectors, that is $\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{W}} \coloneqq \mathbf{x}'\mathbf{W}\mathbf{y} = (\mathbf{L}\mathbf{x})'(\mathbf{L}\mathbf{y})$ where $\mathbf{W} = \mathbf{L}'\mathbf{L}$ for some triangular matrix given via Cholesky decomposition.

Definition 16. (Angle) The angle θ between two vectors $\mathbf{x},\mathbf{y}$ defined in an inner product space is $cos\theta = \frac{\langle\mathbf{x},\mathbf{y}\rangle}{\sqrt{\langle\mathbf{x},\mathbf{x}\rangle\cdot\langle\mathbf{y},\mathbf{y}\rangle}}$.

We can only define angle with the notion of an inner product, later we know that we have quasi-inner product when the matrix $\mathbf{W}$ is semi-positive definite instead of positive definite. Although allowing such a gap might be convenient in fitting some WLS/GLS model, we do not have a angle interpretation of taking quasi-inner product.

With the notion of angle, we can lay out the frequency vectors of two documents' dictionary, the angle inbetween them can be taken as a measure of similarity of two documents.

Only in an inner product space can we talked about orthogonality and hence have the Gram-Schmidt orthogonalization which leads to the rank and QR decomposition we are to discuss.

3. The Inverse

3.1. **Inverse formula.**

Theorem 17. *For a partitioned matrix*

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{pmatrix}$$

, its inverse is given by

$$\begin{pmatrix} \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} & \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21}\right)^{-1} & -\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}\mathbf{A}_{21}\right)^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}\mathbf{A}_{12}\right)^{-1} \end{pmatrix}$$

3.2. **Multivariate normal distribution.** One thing about multivariate normal distribution is that we only call the distribution which is associated with Lebesgue measure and with the p.d.f.

$$\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right]$$

as a p-dimensional multivariate normal distribution. If we change the measure arbitrarily, then it is no longer called a "multivariate normal" distributions. What is more, if the Σ has rank-deficiency, than we do not have p.d.f. of the form above, we then call it a *degenerate multivariate normal distribution*, which can not serve as an example that two random variables' joint distribution is no longer normal. [Anderson] proposed an equivalent definition avoided all these complications.

Another thing to mention about the integration of p.d.f. is that using the Lebesgue measure, the low dimensional sets have measure zero in the high dimensional space.

The following lemma applies to not only multivariate normal random vectors but virtually all random vectors.

Lemma 18. $\mathbb{E}(\beta'\mathbf{X}) = \beta'\mathbb{E}(\mathbf{X}), Var(\beta'\mathbf{X}) = \beta' Var(\mathbf{X})\beta$. *Moreover,* $\mathbb{E}(\mathbf{AX}) = \mathbf{A}\mathbb{E}(\mathbf{X}), Var(\mathbf{AX}) = \mathbf{A}Var(\mathbf{X})\mathbf{A}'$

Now statisticians are interested in extending the distribution from finite variate case to infinite variate case, one of the examples is the Brownian motion. Another branch of statistics is thriving in studying some nice properties if some sort of sparsity is admitted. This branch is the so-called high-dimensional statistics.

3.3. **A Bayesian regression model.**[6] We can construct a hierarchical model by using the conditional multivariate normal distribution. Let $\beta \sim N(\beta_0, \Sigma_\beta); \mathbf{Y}|\beta \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Consider the joint distribution which is again multivariate normal $\mathbf{W} = \begin{pmatrix} \mathbf{Y}|\beta \\ \beta \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}\beta \\ \beta_0 \end{pmatrix}, \begin{pmatrix} \sigma^2\mathbf{I} & \\ & \Sigma_\beta \end{pmatrix}\right)$. Use the following formula for conditional multivariate normal distribution which can be derived by Inverse formula for partitioned matrix or by conditional moments directly:

Lemma 19. *If* $\begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{pmatrix}\right)$, *then*

$$\mathbf{X_1}|\mathbf{x_2} \sim N\left(\mu_1 + \mathbf{\Sigma_{12}\Sigma_{22}}^{-1}\left[\mathbf{x_2} - \mu_2\right], \mathbf{\Sigma_{11}} - \mathbf{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}\right)$$

Thus, if we know the sample moments, we can update the original moment of the distribution of β.

3.4. **Generalized inverse.**

Theorem 20. *(G-inverse of partitioned matrices) If* $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ *with* $\mathbf{B}_{11}$ *the maximal nonsingular principal submatrix, then*

$$\mathbf{B}^- = \begin{pmatrix} \mathbf{B}_{11}^{-1} - \mathbf{X}\mathbf{B}_{21}\mathbf{B}_{11}^{-1} - \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{Y} - \mathbf{B}_{11}^{-1}\mathbf{B}_{12}\mathbf{Z}\mathbf{B}_{21}\mathbf{B}_{11}^{-1} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$$

with $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ *being arbitrary matrices of appropriate dimensions.*

Theorem 21. *(Theorem 9.2.7 in* [Harville]*) Let* $\mathbf{G}$ *be any particular g-inverse to* $\mathbf{A}_{m\times n}$ *matrix, then any* $n \times m$ *matrix* $\mathbf{G}^*$ *is g-inverse to* $\mathbf{A}$ *iff*
$\mathbf{G}^* = \mathbf{G} + \mathbf{Z} - \mathbf{GAZAG}$ *for some matrix* $\mathbf{Z}_{n\times m}$ *iff*
$\mathbf{G}^* = \mathbf{G} + (\mathbf{I} - \mathbf{GA})\mathbf{T} + \mathbf{S}(\mathbf{I} - \mathbf{AG})$ *for some matrices* $\mathbf{T}_{n\times m}, \mathbf{S}_{n\times m}$.

The fact that any matrix has a g-inverse can be verified by constructing a specific g-inverse using left/right inverses. However, the g-inverse constructed in the following way is only *reflexive g-inverse* which does not cover all possibilities described in the Theorem 9.2.7 [Harville]. Interested readers may refer to [Rao&Mitra].

Theorem 22. *(Theorem 9.2.9 in* [Harville]*) If a matrix* $\mathbf{A}$ *has full column rank, then*
(a) $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ *is a left inverse to* $\mathbf{A}$.
(b) $\mathbf{A}'(\mathbf{A}'\mathbf{A})^{-1}$ *is a right inverse to* $\mathbf{A}$.

Lemma 23. *(Lemma 9.3.5 in* [Harville]*) For two matrices* $\mathbf{A}, \mathbf{B}$,
(a) $\mathcal{C}(\mathbf{A}) = \mathcal{C}(\mathbf{B}) \Leftrightarrow \mathbf{B} = \mathbf{A}\mathbf{A}^-\mathbf{B}$
(a) $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{C}) \Leftrightarrow \mathbf{C} = \mathbf{C}\mathbf{A}^-\mathbf{A}$

[6][Ghosh et.al]

4. Decompositions Theorems

4.1. Decompositions related to ranks.

Theorem 24. *(Rank decomposition)Any matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \mathbf{CF}$ *where* $rank(\mathbf{C}) = rank(\mathbf{F}) = rank(\mathbf{A})$.

Theorem 25. *(QR decomposition)Any matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \mathbf{QR}$ *where* $\mathbf{Q'Q} = \mathbf{I}$, $\mathbf{R}$ *is an upper triangular matrix. Further,* $\mathbf{R}$ *can be chosen such that its diagonal are all 1's.*

QR factorization can be derived from the Gram-Schmidt orthogonalization procedure, it is also widely used in many packages for obtaining an inverse matrix.

4.2. Decompositions related to eigenvalues.

Theorem 26. *(Principle Axis Theorem)Let* $\mathbf{S}$ *be a symmetric* $p \times p$ *matrix. Let* $q(t) = |\mathbf{S} - t\mathbf{I}|$. *Then there are* p *real roots to the equation* $q(t) = 0$. *Then there exists an orthogonal matrix* $\mathbf{\Gamma}$ *such that* $\mathbf{S} = \Gamma\mathbf{D}\Gamma', \mathbf{D} = diag(t_1, \cdots, t_p), t_1 = sup_{\mathbf{X} \in \mathbb{R}^p \backslash \mathbf{0}} \frac{\mathbf{X'SX}}{\|\mathbf{X}\|^2}$.

This theorem is also known as the spectral decomposition theorem for symmetric operators. This theorem can be used for rotating the axis of a distribution. For example we can always rotate the multivariate normal distribution to yield a diagonal covariance matrix which means the components of this random vector are mutually independent. This theorem is also the basis for principal component analysis. The matrix Γ is obtained by applying Gram-Schmidt procedure to a specific family of eigenvectors as columns of the matrix. Another form of this theorem can be stated as follows:

Theorem 27. *(Spectral decomposition)Any symmetric matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \sum \mathbf{U\Lambda U'}$ *where* $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_{rank\mathbf{A}})$ *be the eigenvalues of the matrix* $\mathbf{A}$ *and* $\mathbf{U'U} = \mathbf{I}$.

Corollary 28. *(Spectral decomposition for p.s.d matrix)Any positive semi-definite symmetric matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \sum \mathbf{U\Lambda U'}$ *where* $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_{rank\mathbf{A}}), \lambda_i \geq 0$ *be the eigenvalues of the matrix* $\mathbf{A}$ *and* $\mathbf{U'U} = \mathbf{I}$.

Theorem 29. *(Projection decomposition)Let* $\mathbf{A}_i$ *be a* $m \times p_i$ *matrix of rank* $r_i, i = 1, \cdots, k$, *if* $\sum_i r_i = m$ *then the followings are equivalent:*

(a) $\mathbf{A}_i^* \mathbf{A}_j = \mathbf{0}, \forall i \neq j$

(b) $\mathbf{I} = \sum_{i=1}^k \mathbf{A}_i \left(\mathbf{A}_i^* \mathbf{A}_i\right)^- \mathbf{A}_i^*$

Theorem 30. *(SVD decomposition)Any matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \mathbf{U\Sigma V'}$ *where* $\mathbf{\Sigma} = diag(\lambda_1, \cdots, \lambda_{rank\mathbf{A}})$ *be the eigenvalues of the matrix* $\mathbf{A'A}$ *and* $\mathbf{U'U} = \mathbf{I}$, $\mathbf{V'V} = \mathbf{I}$.

4.3. **Decompositions related to bilinear forms.**

Theorem 31. *(LDU decomposition)Any positive semi-definite matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \mathbf{LDU}$ *where* $\mathbf{L}$ *is a unit lower triangular matrix;* $\mathbf{U}$ *is a unit upper triangular matrix;* $\mathbf{D}$ *is a diagonal matrix whose diagonal entries are non-negative.*

Any positive definite matrix $\mathbf{A}$ admits following *unique* decomposition $\mathbf{A} = \mathbf{LDU}$ where $\mathbf{L}$ is a unit lower triangular matrix; $\mathbf{U}$ is a unit upper triangular matrix; $\mathbf{D}$ is a diagonal matrix whose diagonal entries are positive.

Theorem 32. *(U'DU decomposition)Any symmetric positive semi-definite matrix* $\mathbf{A}$ *admits following decomposition* $\mathbf{A} = \mathbf{U}'\mathbf{DU}$ *where* $\mathbf{U}$ *is a unit upper triangular matrix;* $\mathbf{D}$ *is a diagonal matrix whose diagonal entries are non-negative.*

Any symmetric positive definite matrix $\mathbf{A}$ admits following *unique* decomposition $\mathbf{A} = \mathbf{U}'\mathbf{DU}$ where $\mathbf{U}$ is a unit upper triangular matrix; $\mathbf{D}$ is a diagonal matrix whose diagonal entries are positive.

Corollary 33. *(Cholesky decomposition for p.d matrix)Any symmetric positive definite matrix* $\mathbf{A}$ *admits following* unique *decomposition* $\mathbf{A} = \mathbf{T}'\mathbf{T}$ *where* $\mathbf{T}$ *is a unit upper triangular matrix whose diagonal entries are positive.*

5. Regression and Projection

5.1. **The hat matrix.** The hat matrix is the projection matrix onto the column space of a design matrix $\mathbf{X}$, which is to say, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = P_{\mathbf{X}}$. With this notion in head, we could simply drop those redundant columns in the design matrix as long as the column space of design matrix is not altered. This dropping can also be regarded as that we are reducing the corresponding normal equation system into an equivalent system by doing column reduction. See [Strang] Chap.1.

Projections of a certain subspace is unique, but different inner product defined on the same space might lead to different projections. We can also review how Gram-Schmidt orthogonalization leads to an orthonormal basis and regard projections are simply coordinate mapping onto any of such an orthonormal basis of the space it projects onto.

The model we are to consider in this section is $\mathbf{Y} = \mathbf{X}\beta + \epsilon, \epsilon \sim N(\mathbf{0}, \Sigma)$, we can view this model as $\mathbf{W} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\mathbf{1}} \\ \mu_{\mathbf{2}} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{pmatrix} \right)$, our goal is to fit the model using the observation, or simply we can say our goal is to find the conditional distribution $\mathbf{Y}|\mathbf{X}$. Using the multivariate normal distribution result in previous sections, we see that

$$\mathbf{Y}|\mathbf{X} \sim N(\mu_{\mathbf{1|2}}, \mathbf{\Sigma_{1|2}}) = N\left(\mu_{\mathbf{2}} + \mathbf{\Sigma_{21}}\mathbf{\Sigma_{11}^{-1}}(\mathbf{x} - \mu_{\mathbf{1}}), \mathbf{\Sigma_{22}} - \mathbf{\Sigma_{21}}\mathbf{\Sigma_{11}}\mathbf{\Sigma_{12}}\right)$$

. From this conditional distribution it is not hard to see that the mean function and the variance function are exactly linear in the observation/regressors.

Using the iterated expectation/variance formula, it is not hard to see $\mathbb{E}\mathbf{Y} = \mathbf{X}\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, Var\mathbf{Y} = \sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma_{22}}\mathbf{X}'$

Lemma 34. *A matrix* $\mathbf{V}$ *is a covariance matrix iff it is a symmetric positive semi-definite matrix.*

5.2. **Sequential model and Successive projections. Proposition.** The $\|\mathbf{y}-P_{\mathbf{X}}\mathbf{y}\|$ is the minimum of the function $\|\mathbf{y}-\mathbf{w}\|$ in variable $\mathbf{w}$. That i to say it is a least square solution to the corresponding model.

In the sequential model, we usually relate a series of models with a series of projections. The testing sum of squares we encountered in [Weisberg] can be regarded as the norm of projected image onto certain subspaces. Think about the easiest case when projection splits the space into two *direct summands,* one is the fitted value space and the other is the space *where the residuals live in.* These two direct summand spaces are mutually each other's orthogonal complement.

One thing to remember is that the sum of squares are not directly the norm of some projected image yet the difference between two norms.

While dropping regressors might be considered as shrinking the base space of a projection, adding a new regressor is extending the base space of a projection. $P_{[\mathbf{X_1}]} \to P_{[\mathbf{X_1}\mathbf{X_2}]} \to \cdots P_{[\mathbf{X_1}\cdots\mathbf{X_n}]}$

	OLS	WLS	GLS
Covariance matrix	$\sigma^2\mathbf{I}$	$\mathbf{D}^{-1}$	$\mathbf{\Sigma}$
Projection	$\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$	$\mathbf{X}\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}$	$\mathbf{X}\left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{\Sigma}^{-1}$
Parameter estimate	$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$	$\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{Y}$	$\left(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{Y}$

The *twicing* technique on the smoothers are usually used to fasten the fitting procedure of WLS/GLS, they are also used for estimating the weights and covariance matrices in WLS/GLS, see [Weisberg] Chap.4. Not correctly assign such a weight or covariance matrix will generally cause wider confident band although the point estimators themselves are still unbiased. This kind of model WLS/GLS usually occurs when the data is from a time series or nested setup.

References

[Anderson] Anderson, T. W. "Wiley publications in statistics. An introduction to multivariate statistical analysis. Hoboken." (1958).

[Ghosh et.al] Ghosh, Jayanta K., Mohan Delampady, and Tapas Samanta. An introduction to Bayesian analysis: theory and methods. Springer Science & Business Media, 2007.

[Rao&Mitra] Rao, C. Radhakrishna, and Sujit Kumar Mitra. "Generalized inverse of a matrix and its applications." Proceedings of the sixth Berkeley symposium on mathematical statistics and probability. Vol. 1. 1972.

[Ando] Ando, Tomohiro. Bayesian model selection and statistical modeling. CRC Press, 2010.

[Strang] Strang, G. (1988) Linear Algebra and Its Applications, 3rd ed., Harcourt Brace Jovanovich, San Diego,

[Weisberg] Weisberg, Sanford. Applied Linear Regression. John Wiley & Sons, 2013.
[Harville] Harville, D. A. "Matrix Algebra from a Statistician's Perspective. 1997." Inc., Springer-Verlag New York.