Linear regression with random projections (Gaussian weak features model)

Belkin, Hsu, Xu (2019) arXiv: 1903.07571
Breiman, Freedman (1983) JASA vol 78, 361
Lafon, Thomas (2021) arXiv: 2403.10459

Consider a data set $S = \{(\vec{x}^k, y^k)\}_{k=1}^{n}$ with each pair $(\vec{x}^k, y^k) \in \mathbb{R}^{d+1}$ sampled i.i.d from a distribution $\mathcal{D}(\vec{x}, y)$. Assume:

$$\mathcal{D}(\vec{x}) = \mathcal{N}(\vec{x} \,|\, \vec{0}, \mathbb{I}_D)$$

$\mathcal{D}(y|\vec{x})$: modeled by the linear function

$$y = \vec{\beta}^T \vec{x} + \eta \epsilon \, ; \quad \epsilon \sim \mathcal{N}(\epsilon|0,1)$$

$$\eta > 0$$

"ground-truth vector" : $\vec{\beta} \in \mathbb{R}^d$

Let:

$$X \equiv [\vec{x}^1 | \ldots | \vec{x}^n]^T \in \mathbb{R}^{n \times d}$$

$$\vec{y} \equiv [y^1 \, y^2 \ldots y^n]^T \in \mathbb{R}^n$$

Least-squares estimation

In usual least-squares estimation, one searches a minimizer $\hat{\vec{\beta}} \in \mathbb{R}^d$ for the empiral loss quadratic loss:

$$\hat{\vec{\beta}} = \underset{\vec{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \; \mathcal{L}_S(S; \vec{\beta})$$

$$= \underset{\vec{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \; \underbrace{\sum_{k=1}^{n} \left(y^k - \vec{\beta}^T \vec{x}^k\right)^2}_{\text{squared loss function}}$$

$$= \underset{\vec{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \; (\vec{y} - X\vec{\beta})^T (\vec{y} - X\vec{\beta})$$

$$= \underset{\vec{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \; \| \vec{y} - X\vec{\beta} \|^2$$

In the exercises we also discuss a particular kind of regularized least squares.

linear regression with (random) projections

We will fit a linear model to the data using only a subset:

$$A \subseteq [d] \equiv \{1, \dots, d\} \text{ of } p \equiv |A| \text{ variables}$$

For any $\vec{v} \in \mathbb{R}^d$ we use

$$\vec{v}_A \equiv [v_j : j \in A]^T$$

to denote its $|A|$-dimensional subvector of entries from A. Also denote

$$\vec{X}_A \equiv [\vec{x}_A^1 | \dots | \vec{x}_A^n]^T \in \mathbb{R}^{n \times |A|}$$

For $A \subseteq [d]$, its complement is denoted by $A^c \equiv [d] \setminus A$.

The regression coefficients $\hat{\vec{\beta}} \in \mathbb{R}^d$ are fitted with

$$\boxed{\begin{aligned} \hat{\vec{\beta}}_A &= X_A^{\dagger} \vec{y} \\ \hat{\vec{\beta}}_{A^c} &= \vec{0} \end{aligned}}$$

- Solution for the least-squares problem $X_A^T X_A \vec{\beta}_A = X_A^T \vec{y}$ for $\vec{\beta}_A$
- $\vec{\beta}_A^c$ forced to all-zeros

$\dagger$: denotes the Moore-Penrose inverse

Definition : Moore - Penrose inverse

For $M \in \mathbb{R}^{n \times s}$, a Moore-Penrose or pseudo inverse is defined as the matrix $M^{\dagger} \in \mathbb{R}^{s \times n}$ satisfying all the three criteria :

1) $MM^\dagger$ need not be the identity matrix, but it maps all the column vectors of M to themselves:

$$M M M^\dagger = M$$

2) $M^\dagger$ acts like a weak inverse:

$$M^\dagger M M^\dagger = M^\dagger$$

3) $MM^\dagger$ and $M^\dagger M$ are symmetric:

$$(MM^\dagger)^T = MM^\dagger \quad ; \quad (M^\dagger M)^T = M^\dagger M$$

Note that $MM^\dagger$ and $M^\dagger M$ are orthogonal projection operators, as follows from

$$(MM^\dagger)^2 = MM^\dagger \quad ; \quad (M^\dagger M)^2 = M^\dagger M$$

Test risk / generalization error

Once the estimator $\hat{\vec{\beta}} \in \mathbb{R}^d$ is computed, its quality over a new sample pair:

$$\left(\vec{x}^{\,new}, y^{new}\right) \sim \mathcal{D}$$

same distribution that generated the training data

can be measured by:

$$\ell\left(\vec{x}^{\,new}, y^{new}; \hat{\vec{\beta}}\right)$$

prediction error

Conditional mean prediction risk

$$\boxed{\mathcal{E}\left(X, \vec{y}; \hat{\vec{\beta}}\right) \equiv \mathbb{E}_{\vec{x}^{\,new}, y^{new} \mid X, \vec{y}}\left[\ell\left(\vec{x}^{\,new}, y^{new}; \hat{\vec{\beta}}\right)\right]}$$

The test risk is then defined by:

$$\boxed{\mathcal{R}(\vec{\beta}) \equiv \mathbb{E}_{X,\vec{y}}\left[\mathcal{E}(X,\vec{y};\hat{\vec{\beta}})\right]}$$

↳ expectation of conditional prediction error over the data distribution
a.k.a. population risk

Theorem

Assume $\vec{x} \sim \mathcal{N}(\vec{x}\,|\,\vec{0}, \mathbb{I}_d)$; $\epsilon \sim \mathcal{N}(\epsilon|0,1)$ independent of $\vec{x}$ and $y = \vec{\beta}^T\vec{x} + \eta\epsilon$ for some $\vec{\beta} \in \mathbb{R}^d$ and $\eta > 0$. Pick any $p \in \{0,\dots,d\}$ and $A \subseteq [d]$ with $|A| = p$.

For the squared loss:

$$\ell(\vec{x}, y; \hat{\vec{\beta}}) = (y - \hat{\vec{\beta}}^T\vec{x})^2$$

with $\hat{\vec{\beta}}_A = X_A^{\dagger}\vec{y}$ and $\vec{\beta}_{A_c} = \vec{0}$, the test risk $\mathcal{R}_A(\vec{\beta})$ of $\vec{\beta}$ for a given A is:

$$\mathcal{R}_A(\vec{\beta}) =$$

$$\begin{cases} \left(\|\vec{\beta}_{A^c}\|^2 + y^2\right)\left(1 + \dfrac{p}{n-p-1}\right) ; & \underline{\text{if } p \leqslant n-2} \\ +\infty \; ; & \underline{\text{if } n-1 \leqslant p \leqslant n+1} \\ \|\vec{\beta}_A\|^2\left(1 - \dfrac{n}{p}\right) + \left(\|\vec{\beta}_{A^c}\|^2 + y^2\right)\left(1 + \dfrac{n}{p-n-1}\right) ; & \underline{\text{if } p \geqslant n+2} \end{cases}$$

Corollary

Let A be a uniformly random subset of $[d]$ of cardinality p. In the setting of the theorem above, we have:

$$\mathcal{R}(\vec{\beta}) = \mathbb{E}_A\left[\mathcal{R}_A(\vec{\beta})\right] =$$

$$\begin{cases} \left(\left(1-\frac{p}{d}\right)\|\vec{\beta}\|^2+\eta^2\right)\left(1+\frac{p}{n-p-1}\right); & \underline{\text{if } p \leqslant n-2} \\ \|\vec{\beta}\|^2\left[1-\frac{n}{d}\left(2-\frac{d-n-1}{p-n-1}\right)\right] & \\ +\eta^2\left(1+\frac{n}{p-n-1}\right); & \underline{\text{if } p \geqslant n+2} \end{cases}$$

Proof of the corollary

Since A is a uniformly random subset of $[d]$ of cardinality p:

$$\mathbb{E}_A\left[\|\vec{\beta}_A\|^2\right] = \frac{p}{d}\|\vec{\beta}\|^2$$

$$\mathbb{E}_{A^c}\left[\|\vec{\beta}_{A^c}\|^2\right] = \left(1-\frac{p}{d}\right)\|\vec{\beta}\|^2$$

Plugging into Theorem 1 completes the proof.

Sketching the plot for the corollary

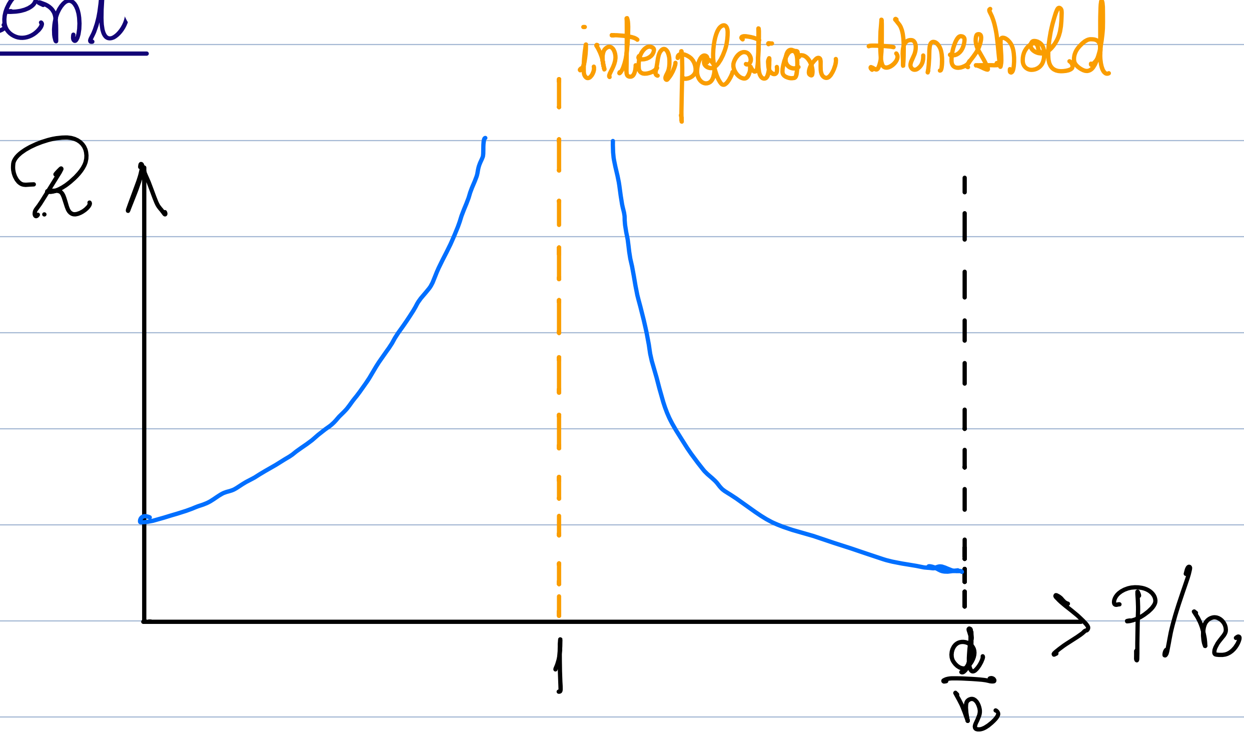
Assume $d > n+1$:

a) The risk first increases with p up to the "interpolation threshold" $p = n$, after which decreases with p.

b) If

$$\frac{\|\vec{\beta}\|^2}{\sigma^2} > \frac{d}{d-n-1},$$

the smallest test risk is achieved at $p = d$. It is smaller than any $p \leq n$: double descent



Proof of the theorem

Let us consider the conditional mean squared risk:

$$\mathcal{E}(X,\vec{y};\hat{\vec{\beta}}) = \mathbb{E}_{\vec{x}^{new},y^{new}}\left[\left(y^{new}-\hat{\vec{\beta}}^T\vec{x}^{new}\right)^2\right] =$$

$$= \mathbb{E}_{\vec{x}^{new},y^{new}}\left[\left((\vec{\beta}-\hat{\vec{\beta}})^T\vec{x}^{new}+\eta\epsilon^{new}\right)^2\right] =$$

$$= (\vec{\beta}-\hat{\vec{\beta}})^T\underbrace{\mathbb{E}_{\vec{x}^{new}}\left[\vec{x}^{new}\vec{x}^{new\,T}\right]}_{\mathbb{I}_d}(\vec{\beta}-\hat{\vec{\beta}})$$

$$+\eta^2\underbrace{\mathbb{E}_{\epsilon^{new}}\left[(\epsilon^{new})^2\right]}_{=1}+2\eta(\vec{\beta}-\hat{\vec{\beta}})^T\underbrace{\mathbb{E}_{\vec{x}^{new}}\left[\vec{x}^{new}\right]}_{=0}\underbrace{\mathbb{E}_{\epsilon^{new}}\left[\epsilon^{new}\right]}_{=0}$$

$$= \sum_{j\in A}(\beta_j-\hat{\beta}_j)^2+\sum_{\ell\in A^c}(\beta_\ell-\hat{\beta}_\ell)^2+\eta^2$$

Since $\hat{\vec{\beta}}_{A_c} = \vec{0}$ we have:

$$\boxed{\mathcal{E}(X,\vec{y};\hat{\vec{\beta}}) = \eta^2+\|\vec{\beta}_{A_c}\|^2+\|\vec{\beta}_A-\hat{\vec{\beta}}_A\|^2}$$

a) Case $p \leqslant n$ Breiman, Freedman 1983

From the least-squares solution on A we have (see exercises):

$$\hat{\vec{\beta}}_A = (X_A^T X_A)^{-1} X_A^T \vec{y}$$

Note that:

$$\text{rank}\underbrace{(X_A^T X_A)}_{} \leqslant \min(n, p) = p$$

Wishart matrix with n degrees of freedom and covariance matrix $\mathbb{I}_p$

↳ for $p \leqslant n$, it is full rank with high probability.

Let $\vec{\eta} \equiv \underbrace{\vec{y} - X_A \vec{\beta}_A}_{}$ and write:

observe that this is a vector in A^c: $\vec{\eta} = X_A \vec{\beta}_A + X_{A^c} \vec{\beta}_{A^c} + \eta \vec{\epsilon} - X_A \vec{\beta}_A$

$$\vec{\beta}_A - \hat{\vec{\beta}}_A = \vec{\beta}_A - (X_A^T X_A)^{-1} X_A^T (\vec{\eta} + X_A \vec{\beta}_A)$$

$$= \vec{\beta}_A - \underline{(X_A^T X_A)^{-1}} \; \underline{X_A^T X_A} \vec{\beta}_A$$

$$- (X_A^T X_A)^{-1} X_A^T \vec{\eta}$$

$$= - (X_A^T X_A)^{-1} X_A^T \vec{\eta}$$

Now the square of the l_2-norm:

$$\| \vec{\beta}_A - \hat{\vec{\beta}}_A \|^2 = \| (X_A^T X_A)^{-1} X_A^T \vec{\eta} \|^2 =$$

$$= \left((X_A^T X_A)^{-1} X_A^T \vec{\eta} \right)^T \left((X_A^T X_A)^{-1} X_A^T \vec{\eta} \right)$$

$$= \vec{\eta}^T X_A \left((X_A^T X_A)^{-1} \right)^T (X_A^T X_A)^{-1} X_A^T \vec{\eta} =$$

$$= \vec{\eta}^T X_A (X_A^T X_A)^{-2} X_A^T \vec{\eta}$$

Let S_A be the unique positive definite square root of $X_A^T X_A$ and define

$$\Psi_A = X_A S_A^{-1} \in \mathbb{R}^{n \times p}$$

Then Ψ_A is orthonormal:

$$X_A^T X_A = (\Psi_A S_A)^T (\Psi_A S_A) = S_A^T \underbrace{\Psi_A^T \Psi_A}_{= \mathbb{I}_p} S_A$$

$$= S_A^2$$

Trace trick

For a matrix $A \in \mathbb{R}^{p \times p}$ and a vector $\vec{u} \in \mathbb{R}^d$, the quantity $\vec{u}^T A \vec{u}$ is a real number and can be thought as a 1×1 matrix. Using the cyclic property of trace:

$$\vec{u}^T A \vec{u} = \mathrm{Tr}\{\vec{u}^T A \vec{u}\} = \mathrm{Tr}\{\vec{u}\vec{u}^T A\} = \mathrm{Tr}\{A \vec{u}\vec{u}^T\}$$

Then:

$$\|\vec{\beta}_A - \hat{\vec{\beta}}_A\|^2 = \vec{\eta}^T \Psi_A S_A \, S_A^{-4} \, S_A^T \Psi_A^T \vec{\eta}$$

$$= \vec{\eta}^T \Psi_A \, S_A^{-2} \, \Psi_A^T \vec{\eta} =$$ trace trick

$$= \vec{v}^T (X_A^T X_A)^{-1} \vec{v}$$

$$= \mathrm{Tr}\{\vec{v}^T (X_A^T X_A)^{-1} \vec{v}\} =$$

$$= \mathrm{Tr}\{(X_A^T X_A)^{-1} \vec{v}\,\vec{v}^T\}$$

where

$$\vec{v} \equiv \Psi_A^T \vec{\eta} \in \mathbb{R}^p$$

Observe that:

$$\vec{v} = \Psi_A^T (\vec{y} - X\vec{\beta}_A) = \Psi_A^T (X_{A^c}\vec{\beta}_{A^c} + \mu\vec{\epsilon})$$

$$\Rightarrow \mathbb{E}_{X_{A^c},\vec{\epsilon}}[\vec{v}] = \vec{0}$$

and:

$$\vec{v}\vec{v}^T = \Psi_A^T\left(X_{A^c}\vec{\beta}_{A^c} + \mu\vec{\epsilon}\right)\left(X_{A^c}\vec{\beta}_{A^c} + \mu\vec{\epsilon}\right)^T\Psi_A =$$

$$= \Psi_A^T\left(X_{A^c}\vec{\beta}_{A^c} + \mu\vec{\epsilon}\right)\left(\vec{\beta}_{A^c}^T X_{A^c}^T + \mu\vec{\epsilon}^T\right)\Psi_A =$$

$$= \Psi_A^T\Big(X_{A^c}\beta_{A^c}\vec{\beta}_{A^c}^T X_{A^c}^T + \mu X_{A^c}\beta_{A^c}\vec{\epsilon}^T$$

$$+ \mu\vec{\epsilon}\vec{\beta}_{A^c}^T X_{A^c}^T + \mu^2\vec{\epsilon}\vec{\epsilon}^T\Big)\Psi_A$$

$\Rightarrow$

$$\mathbb{E}_{X_{A^c},\vec{y},\vec{\epsilon}}\left[\vec{v}\vec{v}^T\right] =$$

$$= \Psi_A^T\,\mathbb{E}_{X_{A^c}}\left[X_{A^c}\vec{\beta}_{A^c}\left(X_{A^c}\beta_{A^c}\right)^T\right]\Psi_A$$

$$+ \mathbb{E}_{X_{A^c}}\left[\Psi_A^T\,\underbrace{\mathbb{E}_{\vec{\epsilon}}\left[\vec{\epsilon}\vec{\epsilon}^T\right]}_{\mathbb{I}_n}\,\Psi_A\right]$$

and $\Psi_A^T\Psi_A = \mathbb{I}_p$

Observe that:

$$\left(X\vec{\beta}\left(X\vec{\beta}\right)^T\right)_{kk'} = \left(X\vec{\beta}\right)_k\left(X\beta\right)_{k'}$$

$$= \sum_{l=1}^{p} x_{kl}\beta_l \sum_{l'=1}^{p} x_{k'l'}\beta_{l'} \Rightarrow$$

$$\mathbb{E}\left[\left(X\vec{\beta}(X\vec{\beta})^T\right)_{kk'}\right] =$$

$$= \sum_{l=1}^{p} \beta_l^2 \underbrace{\mathbb{E}[x_{kl}x_{k'l}]}_{\sigma_{kk'}} = \|\vec{\beta}\|^2 \sigma_{kk'}$$

Thus, since

$$\mathbb{E}_{X_{A^c}, \vec{\epsilon}}\left[X_{A^c}\vec{\beta}_{A^c}(X_{A^c}\vec{\beta}_{A^c})^T\right] = \|\vec{\beta}_{A^c}\|^2 \mathbb{I}_n$$

we have for $\vec{v}$:

$$\mathbb{E}_{X_{A^c}, \vec{\epsilon}}[\vec{v}] = \vec{0}$$

$$\mathbb{E}_{X_{A^c}, \vec{\epsilon}}[\vec{v}\vec{v}^T] = \left(\|\vec{\beta}_{A^c}\|^2 + \eta^2\right)\mathbb{I}_p$$

Thus, since $\Psi_A^T \Psi_A = \mathbb{I}_p$ we can write:

$$\mathbb{E}_{X,\vec{y},\vec{\epsilon}}\left[\|\vec{\beta}_A-\hat{\vec{\beta}}_A\|^2\right]=\mathbb{E}_{X_A,X_{A^c},\vec{\epsilon}}\left[\|\vec{\beta}_A-\hat{\vec{\beta}}_A\|^2\right]$$

$$=\mathrm{Tr}\left\{\mathbb{E}_{X_A}\left[(X_A^T X_A)^{-1}\right]\mathbb{E}_{X_{A^c},\vec{\epsilon}}\left[\vec{v}\vec{v}^T\right]\right\}$$

$$=\left(\|\vec{\beta}_{A^c}\|^2+\eta^2\right)\mathrm{Tr}\left\{\mathbb{E}_{X_A}\left[(X_A^T X_A)^{-1}\mathbb{I}_p\right\}\right.$$

Observe that $X_A\in\mathbb{R}^{n\times p}$ is a matrix with n independent columns sampled from $\mathcal{N}(\vec{x}_A\,|\,\vec{0};\,\mathbb{1}_p)$. Then

$X_A^T X_A$: Wishart matrix

$$\Rightarrow \mathbb{E}_{\vec{x}_A}\left[X_A^T X_A\right]=n\,\mathbb{I}_p$$

$$\mathbb{E}\left[(X_A^T X_A)_{jk}\right]=\sum_{l=1}^{n}\mathbb{E}\left[x_A^{lj}\,x_A^{lk}\right]=$$

$$= \sum_{\ell=1}^{n} \delta_{jk} = n\,\delta_{jk}$$

Book : Potters and Bouchaud
A first course in Random Matrix Theory

The matrix $(X_A^T X_A)^{-1}$, on the other hand, is an inverse-Wishart matrix, which follows a inverse-Wishart distribution with scale matrix $\mathbb{I}_p$ and n degrees of freedom.

↳ the distribution is well-defined for $n > p+1$

* When the scale matrix is the identity, diagonal elements of $(X_A^T X_A)^{-1}$ follow an inverse χ^2 distribution : inv $-\chi^2_{n-p+1}$

The first moment of a inverse-Wishart distribution with scale matrix $\mathbb{I}_p$ and n degrees of freedom is known to be

$$\mathbb{E}_{X_A}\left[(X_A^T X_A)^{-1}\right] = \frac{\mathbb{I}_p}{n-p-1}$$

which is well-defined <u>only</u> for $n > p+1$

Plugging this result in the expression we had before

$$\mathcal{R}_A(\vec{\beta}) = \left(\|\vec{\beta}_{A_c}\|^2 + \eta^2\right)\left(1 + \frac{p}{n-p-1}\right)$$

if $p \leq n-2$

Inverse-Wishart distribution

For a positive definite $M \in \mathbb{R}^{p \times p}$, the PDF of the inverse Wishart is

$$f_M(M; S, n) = \frac{(\det S)^{n/2}}{2^{pn/2}\,\Gamma_p\left(\frac{n}{2}\right)} (\det M)^{-(p+n+1)/2} e^{-\frac{1}{2}\mathrm{Tr}\{S M^{-1}\}}$$

with $S \in \mathbb{R}^{p \times p}$ positive definite being the scale matrix with, by definition $S_{j\ell} \geq 0$, $\forall j, \ell = 1, \ldots, p$. The number $n > p + 1$ is called the degrees of freedom. The function $\Gamma_p(\cdot)$ is the multivariate gamma function.

The first moment is

$$\mathbb{E}_M[M] = \frac{S}{n - p - 1} \quad ; \quad n > p - 1$$

b) Case $p \geq n$

Remembering that $\hat{\vec{\beta}}_A = X_A^\dagger \vec{y}$ and writing the Moore-Penrose inverse as

$$X_A^\dagger = X_A^T (X_A X_A^T)^\dagger$$

Property of the Moore-Penrose inverse when $X_A X_A^T$ and $X_A^T X_A$ are symmetric

and defining:

$$\vec{\eta} \equiv \vec{y} - X_A \vec{\beta}_A$$

we write:

$$\vec{\beta}_A - \hat{\vec{\beta}}_A = \vec{\beta}_A - X_A^\dagger \vec{y} =$$

$$= \vec{\beta}_A - X_A^T (X_A X_A^T)^\dagger (\vec{\eta} + X_A \vec{\beta}_A)$$

$$= \vec{\beta}_A - X_A^T (X_A X_A^T)^\dagger X_A \vec{\beta}_A$$

$$- X_A^T (X_A X_A^T)^\dagger \vec{\eta}$$

$$= \left(\mathbb{I}_p - X_A^T (X_A X_A^T)^\dagger X_A \right) \vec{\beta}_A$$

$$- X_A^T (X_A X_A^T)^\dagger \vec{\eta}$$

$$= \underline{\underline{\left(\mathbb{I}_p - \mathcal{P}_{X_A} \right) \vec{\beta}_A}} - \underline{\underline{X_A^T (X_A X_A^T)^\dagger \vec{\eta}}}$$

where we have defined the projection matrix onto the row space of X_A:

$$\boxed{\mathcal{P}_{X_A} \equiv X_A^T (X_A X_A^T)^\dagger X_A}$$

$$\Rightarrow \left(\mathbb{I}_p - \mathcal{P}_{X_A} \right) \vec{\beta}_A$$: orthogonal projection of $\vec{\beta}_A$ onto the (kernel) null space of X_A

$$\underline{\underline{X_A^T (X_A X_A^T)^\dagger \vec{\eta}}}$$

$$\Rightarrow \quad X_A^T (X_A X_A^T)^{\dagger} \vec{\eta}$$: vector on the row space of X_A

These two vectors are then orthogonal:

$$\|\vec{\beta}_A - \hat{\vec{\beta}}_A\|^2 =$$

$$= \underline{\underline{\|(\mathbb{I}_p - P_{\vec{X}_A})\vec{\beta}_A\|^2}} + \underline{\underline{\|X_A^T (X_A X_A^T)^{\dagger} \vec{\eta}\|^2}}$$

<u>First term</u>

$$\|(\mathbb{I}_p - P_{\vec{X}_A})\vec{\beta}_A\|^2$$

$$= \vec{\beta}_A^T (\mathbb{I}_p - P_{\vec{X}_A}^T)(\mathbb{I}_p - P_{\vec{X}_A})\vec{\beta}_A =$$

$$= \vec{\beta}_A^T \vec{\beta}_A - \vec{\beta}_A^T P_{X_A} \vec{\beta}_A - \vec{\beta}_A^T P_{X_A}^T \vec{\beta}_A + \vec{\beta}_A^T P_{X_A}^T P_{X_A} \vec{\beta}_A$$

Observe that:

$$\mathcal{P}_{X_A}^{T} = \left(X_A^T (X_A X_A^T)^\dagger X_A\right)^T = X_A^T \left(X_A^T (X_A X_A^T)^\dagger\right)^T =$$

$$= X_A^T \left((X_A X_A^T)^\dagger\right)^T X_A = X_A^T (X_A X_A^T)^\dagger X_A$$

$$= \mathcal{P}_{X_A}$$

Also note that:

$$\mathcal{P}_{X_A}^2 = X_A^T (\underbrace{X_A X_A^T}) ^\dagger X_A X_A^T (X_A X_A^T)^\dagger X_A$$

$n \times n$ full rank matrix with high probability, as $\vec{X}_A$ is full row rank with high probability: has "almost" n linearly independent rows (with high probability)

$\Rightarrow (X_A X_A^T)^\dagger X_A X_A^T \approx \mathbb{I}_n$

$$= X_A^T (X_A X_A^T)^\dagger X_A = \mathcal{P}_{X_A}$$

Then we have:

$$\left\| \left(\mathbb{I}_p - P_{X_A} \right) \vec{\beta}_A \right\|^2 =$$

$$= \vec{\beta}_A^T \vec{\beta}_A - 2 \vec{\beta}_A^T P_{X_A}^T P_{X_A} \vec{\beta}_A + \vec{\beta}_A^T P_{X_A}^T P_{X_A} \vec{\beta}_A$$

$$= \vec{\beta}_A^T \vec{\beta}_A - \left(P_{X_A} \vec{\beta}_A \right)^T \left(P_{X_A} \vec{\beta}_A \right)$$

$\Rightarrow$

$$\left\| \left(\mathbb{I}_p - P_{X_A} \right) \vec{\beta}_A \right\|^2 =$$

$$= \| \vec{\beta}_A \|^2 - \| P_{X_A} \vec{\beta}_A \|^2$$

By rotation symmetry of the standard normal distribution we have:

$$\mathbb{E}_{X, \vec{y}} \left[\| P_{X_A} \vec{\beta}_A \|^2 \right] = \frac{n}{p} \| \vec{\beta}_A \|^2$$

* at the end there is a remark on this

$\Rightarrow$

$$\mathbb{E}_{X,\vec{y}}\left[\left\|\left(\mathbb{I}_p - \mathcal{P}_{X_A}\right)\vec{\beta}_A\right\|^2\right]$$
$$= \left\|\vec{\beta}_A\right\|^2\left(1 - \frac{n}{p}\right)$$

Second term

$$\left\|X_A^T\left(X_A X_A^T\right)^{\dagger}\vec{\eta}\right\|^2 =$$

$$= \mathrm{Tr}\left\{\left(X_A^T\left(X_A X_A^T\right)^{\dagger}\vec{\eta}\right)^T X_A^T\left(X_A X_A^T\right)^{\dagger}\vec{\eta}\right\} =$$

$$= \mathrm{Tr}\left\{\vec{\eta}^T\left(X_A^T\left(X_A X_A^T\right)^{\dagger}\right)^T X_A^T\left(X_A X_A^T\right)^{\dagger}\vec{\eta}\right\} =$$

$$= \mathrm{Tr}\left\{\left(\left(X_A X_A^T\right)^{\dagger}\right)^T X_A\ X_A^T\left(X_A X_A^T\right)^{\dagger}\vec{\eta}\vec{\eta}^T\right\} =$$

$$= \mathrm{Tr}\left\{\left(X_A X_A^T\right)^{\dagger}\underbrace{X_A X_A^T\left(X_A X_A^T\right)^{\dagger}}_{\text{almost surely } \mathbb{I}_n \text{ because}}\vec{\eta}\vec{\eta}^T\right\}$$

$X_A X_A^T \in \mathbb{R}^{n\times n}$ $(n<p)$ is almost surely invertible (with high probability)

$$= \mathrm{Tr}\left\{ (X_A X_A^T)^{\dagger} \vec{\eta}\,\vec{\eta}^T \right\}$$

Remember that:

$$\vec{\eta} = \underbrace{\vec{y}}_{\underline{X_A \vec{\beta}_A} + X_{A^c} \vec{\beta}_{A^c} + \gamma \vec{\epsilon}} - \underline{X_A \vec{\beta}_A}$$

is a vector in A^c:

$$\vec{\eta} = X_{A^c} \vec{\beta}_{A^c} + \gamma \vec{\epsilon} \in \mathbb{R}^n$$

Then since $X_A \vec{\beta}_A$ and $X_{A^c} \vec{\beta}_{A^c} + \gamma \vec{\epsilon}$ are uncorrelated we have

$$\mathbb{E}_{X,\vec{y}}\left\{\| X_A^T (X_A X_A^T)^\dagger \vec{\eta} \|^2\right\} =$$

$$= \operatorname{Tr}\left\{ \mathbb{E}_{X,y}\left[(X_A X_A^T)^\dagger\right] \mathbb{E}_{X,\vec{y}}\left[\vec{\eta}\vec{\eta}^T\right] \right\} =$$

$$= \operatorname{Tr}\left\{ \underline{\underline{\mathbb{E}_{X_A}\left[(X_A X_A^T)^\dagger\right]}} \, \underline{\underline{\mathbb{E}_{X_{A^c}} \mathbb{E}_{\vec{\epsilon}}\left[\vec{\eta}\vec{\eta}^T\right]}} \right\}$$

Note that

$$\vec{\eta}\vec{\eta}^T = (X_{A^c}\vec{\beta}_{A^c} + \mu\vec{\epsilon})(\vec{\beta}_{A^c}^T X_{A^c}^T + \mu\vec{\epsilon}^T)$$

$$= \|\vec{\beta}_{A^c}\|^2 X_{A^c} X_{A^c}^T + \mu X_{A^c}\vec{\beta}_{A^c}\vec{\epsilon}^T$$

$$+ \mu\vec{\epsilon}\vec{\beta}_{A^c}^T X_{A^c}^T + \mu^2 \vec{\epsilon}\vec{\epsilon}^T$$

$$\Rightarrow \underline{\underline{\mathbb{E}_{X,\vec{y}}\mathbb{E}_{\vec{\epsilon}}\left[\vec{\eta}\vec{\eta}^T\right] = (\|\vec{\beta}_{A^c}\|^2 + \mu^2) I_n}}$$

Let us now define :

$$\Pi_A \equiv \left(X_A X_A^T \right)^{\dagger}$$

Observe that $X_A \in \mathbb{R}^{n \times p}$ is a matrix with n independent columns sampled from $\mathcal{N}(\vec{x}_A \mid \vec{0}; \mathbb{1}_p)$. Then

$X_A X_A^T$: Wishart matrix

$$\hookrightarrow \mathbb{E}_{X_A}\left[X_A X_A^T \right] = p \, \mathbb{I}_n$$

$$\mathbb{E}\left[(X_A X_A^T)_{jk} \right] = \sum_{l=1}^{p} \mathbb{E}\left[X_A^{jl} X_A^{kl} \right] =$$

$$= \sum_{l=1}^{p} \delta_{jk} = p \, \delta_{jk} \;//$$

Book : Potters and Bouchaud
A first course in Random Matrix Theory

Inverse - Wishart distribution

For a positive definite $M \in \mathbb{R}^{n \times n}$, the PDF of the inverse Wishart is

$$f_M(M; S, p) = \frac{(\det S)^{p/2}}{2^{pn/2}\, \Gamma_n\left(\frac{p}{2}\right)} (\det M)^{-(p+n+1)/2}\, e^{-\frac{1}{2}\mathrm{Tr}\{S M^{-1}\}}$$

with $S \in \mathbb{R}^{n \times n}$ positive definite being the scale matrix with, by definition $S_{je} \geq 0$, $\forall j, l = 1, \ldots, n$. The number $p > n+1$ is called the degrees of freedom. The function $\Gamma_n(\cdot)$ is the multivariate gamma function.

The first moment is

$$\mathbb{E}_M[M] = \frac{S}{p-n-1} \; ; \quad p > n-1$$

The matrix Π_A, on the other hand, is an inverse-Wishart matrix, which follows a inverse-Wishart distribution with scale matrix $\mathbb{I}_n$ and p degrees of freedom.

↳ the distribution is well-defined for $p > n+1$

* When the scale matrix is the identity, diagonal elements of Π_A follow an inverse χ^2 distribution : inv - χ^2_{p-n+1}

The first moment of a inverse-Wishart distribution with scale matrix $\mathbb{I}_n$ and p degrees of freedom is known to be

$$\mathbb{E}_{X_A}\left[\Pi_A\right] = \frac{\mathbb{I}_n}{p-n-1}$$

which is well-defined only for $p > n+1$.

For $p = n+1$, an extrapolation of the result above makes the first moment infinite, which can be interpreted as the PDF going to zero, as one can write $e^{-\mathrm{Tr}\{\cdot\}}$ in terms of the first moment.

The case $p = n$ can also be interpreted as the expectation going to infinite in order to send the PDF to zero.

↳ It is also consistent with Breiman, Freedman (1983)

Therefore for $p > n+1$:

$$\mathbb{E}_{X, \vec{y}} \left\{ \| X_A^\top (X_A X_A^\top)^\dagger \vec{\eta} \|^2 \right\} =$$

$$\mathrm{Tr} \left\{ \frac{\mathbb{I}_n}{p-n-1} \left(\| \vec{\beta}_{A^c} \|^2 + \eta^2 \right) \mathbb{I}_n \right\}$$

$$= \frac{n}{p-n-1}\left(\|\vec{\beta}_{A^c}\|^2 + \eta^2\right)$$

Summing the first term with the second term we finally obtain:

$$\boxed{\mathcal{R}_A(\vec{\beta}) = \begin{cases} +\infty \; ; \; \text{if } n-1 \leqslant p \leqslant n+1 \\ \|\vec{\beta}_A\|^2\left(1-\frac{n}{p}\right) + \left(\|\vec{\beta}_{A^c}\|^2+\eta^2\right)\left(1+\frac{n}{p-n-1}\right) \\ \qquad \text{if } p \geqslant n+2 \end{cases}}$$

$\blacksquare$

Remark on:

$$\mathbb{E}_{X,\vec{y}}\left[\|P_{X_A}\vec{\beta}_A\|^2\right] = \frac{n}{p}\|\vec{\beta}_A\|^2$$

Observe that

$$\|P_{X_A}\vec{\beta}_A\|^2 = (P_{X_A}\vec{\beta}_A)^T(P_{X_A}\vec{\beta}_A) =$$

$$= \vec{\beta}_A^T P_{X_A}^T P_{X_A}\vec{\beta}_A = \vec{\beta}_A^T P_{X_A}^2\vec{\beta}_A =$$

$$= \vec{\beta}_A^T P_{X_A}\vec{\beta}_A = \mathrm{Tr}\{\vec{\beta}_A^T P_{X_A}\vec{\beta}_A\} =$$

$$= \mathrm{Tr}\{\vec{\beta}_A\vec{\beta}_A^T P_{X_A}\}$$

$$\Rightarrow \mathbb{E}_{X_A}\left[\|P_{X_A}\vec{\beta}_A\|^2\right] =$$

$$= \mathrm{Tr}\{\vec{\beta}_A\vec{\beta}_A^T\,\mathbb{E}_{X_A}[P_{X_A}]\}$$

The matrix P_{X_A} can be viewed as as projection matrix, that projects vectors in $\mathbb{R}^p$ onto the column space of X_A.

Given that the rows of X_A are iid $\mathcal{N}(\vec{x}_A \mid \vec{0}_p, \mathbb{I}_p)$, none of the p directions should be preffered. Then its reasonable to expect:

$$\mathbb{E}_{X_A}\left[\mathcal{P}_{X_A}\right] \propto \mathbb{I}_p$$

isotropic rotational symmetry

Since $\mathcal{P}_{X_A}$ is a projection matrix of rank n, with high probability, its trace (the sum of its eigenvalues, which are either 0 or 1; remember that $\mathcal{P}_{X_A} = \mathcal{P}_{X_A}^2$) should be n, with high probability. Then, its reasonable to expect:

$$\mathrm{Tr}\left\{\mathbb{E}_{X_A}\left[\mathcal{P}_{X_A}\right]\right\} = n$$

Thus assuming that $\mathbb{E}_{X_A}\left[\mathcal{P}_{X_A}\right]$ must be isotropic and proportional to

$\mathbb{I}_p$ the scale factor that ensures $\mathrm{Tr}\{\mathbb{E}_{X_A}[\mathcal{P}_{X_A}]\} = n$ is n/p. Then:

$$\mathbb{E}_{X_A}[\mathcal{P}_{X_A}] = \frac{n}{p}\mathbb{I}_p$$

The projection $\mathcal{P}_{X_A}$ essentially distributes the effect of the n dimensions uniformly (on average) across the p components.

Finally:

$$\mathbb{E}_{X_A}\left[\|\mathcal{P}_{X_A}\vec{\beta}_A\|^2\right] = \mathrm{Tr}\left\{\vec{\beta}_A\vec{\beta}_A^T\,\mathbb{E}_{X_A}[\mathcal{P}_{X_A}]\right\}$$

$$= \mathrm{Tr}\left\{\vec{\beta}_A\vec{\beta}_A^T\,\mathbb{I}_p\frac{n}{p}\right\}$$

$$= \frac{n}{p}\mathrm{Tr}\left\{\vec{\beta}_A\vec{\beta}_A^T\right\}$$

$$= \frac{n}{p}\sum_{j=1}^{p}\beta_{A,j}^2 = \frac{n}{p}\|\vec{\beta}_A\|^2$$