Confidence Sets and Gaussian Correlation Inequality

Tianyu Zhou and Shuyang Shen

November 25, 2016

1 Introduction

Royen[1] recently gave a 'simple' proof of the Gaussian correlation inequality, which states that the product of the Gaussian measure of two symmetric convex sets is no greater than the measure of their product.

In this paper, we shall go over the proof of this theorem as explained by the overviews by Panchenko[2] and Latała and Matlak[3], and consider its application to confidence rectangles of random vectors of normal distributions[4].

2 Gaussian Correlation Inequality

2.1 Introduction

Definitions. A set $S \subset \mathbb{R}^n$ is *symmetric* if for every $x \in S$, we have $-x \in S$.

A set $P \subset \mathbb{R}^n$ is a *symmetric strip* if it is of the form

$$P = \{x \in \mathbb{R}^n : |\langle x, v\rangle| \leq t\}$$

for some $v \in \mathbb{R}^n, t \geq 0$.

Theorem 1 (Gaussian Correlation Inequality). *For any closed symmetric convex sets* K, L *in* $\mathbb{R}^d$ *and any centered Gaussian measure* μ *on* $\mathbb{R}^d$, *we have*

$$\mu(K \cap L) \geq \mu(K)\mu(L). \tag{1}$$

Note that any symmetric closed convex set is a countable intersection of symmetric strips by taking the symmetric strips along the tangent line of each differentiable rational point on the boundary, so it is enough to prove that the theorem holds when K, L are symmetric strips,

$$K = \{x \in \mathbb{R}^d : |\langle x, v_i\rangle| \leq t_i \forall 1 \leq i \leq n_1\}, L = \{x \in \mathbb{R}^d : |\langle x, v_i\rangle| \leq t_i \forall n_1 + 1 \leq i \leq n_1 + n_2\}.$$

Take $n = n_1 + n_2$, $X_i = \langle v_i, G\rangle$, then we can obtain an equivalent form of Gaussian Correlation Inequality.

Theorem 2 (Reworded Gaussian Correlation Inequality). *For any* $t_i > 0$*, we have*

$$\mathbb{P}(|X_1| \le t_1, \dots, |X_n| \le t_n) \ge \mathbb{P}(|X_1| \le t_1, \dots, |X_{n_1}| \le t_{n_1})\mathbb{P}(|X_{n_1+1}| \le t_{n_1+1}, \dots, |X_n| \le t_n). \tag{2}$$

2.2 Proof of the Theorem

We shall now go through the proof of the theorem. Note that the proof refers to several lemmas, which are stated and proven in the Auxiliary Lemmas section to come.

First we need a few definitions on the concepts used.

Definitions. A matrix $A \in M_{n\times n}$ is *positively defined* if for any $x \in \mathbb{R}^n$, $x^T A x \ge 0$. It is *strictly positively defined* if equality holds only at 0.

For $A, B \in M_{n\times n}$, we write $A \ge B$ if $A - B$ is positively defined.

Note that covariance matrices are positively defined, as we have

$$x^T C x = \mathbb{E}x^T(X - \mathbb{E}X)(X - \mathbb{E}X)x = \mathbb{E}((X - \mathbb{E}X)x)^T(X - \mathbb{E}X)x \ge 0.$$

Step I: Use the idea of interpolation for Gaussian distributions to transfer our problem to an equivalent.

Without loss of generality, we will assume that the covariance matrix C of X is strictly positively defined, where C can be written as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where C_{ij} is the $n_i \times n_j$ matrix. Define

$$C(\tau) := \begin{pmatrix} C_{11} & \tau C_{12} \\ \tau C_{21} & C_{22} \end{pmatrix}, \quad 0 \le \tau \le 1.$$

And set $Z_i(\tau) := \frac{1}{2}X_i(\tau)^2, 1 \le i \le n$, where $X(\tau) \sim \mathcal{N}(0, C(\tau))$.

Notice that $C(\tau)$ is also strictly positively defined, which means it's a well-defined covariance matrix.

$$z^T C(\tau) z = \begin{pmatrix} a^T & b^T \end{pmatrix} \begin{pmatrix} C_{11} & \tau C_{12} \\ \tau C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \left(a^T C_{11} a + b^T C_{22} b\right) + \tau\left(b^T C_{21} a + a^T C_{12} b\right)$$

Consider about this formula: the first term is non-negative and only the second term could be negative. If the second term is non-negative, then clearly, $z^T C(\tau) z$ is strictly positive. If the second term is negative, then

$$z^T C(\tau) z \ge \left(a^T C_{11} a + b^T C_{22} b\right) + \left(b^T C_{21} a + a^T C_{12} b\right) = \begin{pmatrix} a^T & b^T \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = z^T C z > 0,$$

since τ belongs to [0,1].

Then we can restate the theorem as

$$\mathbb{P}(Z_1(1) \le s_1, \dots, Z_n(1) \le s_n) \ge \mathbb{P}(Z_1(0) \le s_1, \dots, Z_n(0) \le s_n), \tag{3}$$

where $s_i = \frac{1}{2}t_i^2$. To prove the theorem, it is enough to show that the function

$$\tau \mapsto \mathbb{P}(Z_1(\tau) \le s_1, \dots, Z_n(\tau) \le s_n) \tag{4}$$

is nondecreasing on [0, 1].

Let $f(x, \tau)$ be the density of the random vector $Z(\tau)$ and and $K = [0, s_1] \times \cdots \times [0, s_n]$. We get the derivative of $f(x, \tau)$ w.r.t. τ

$$\frac{\partial}{\partial\tau}\mathbb{P}(Z_1(\tau) \le s_1, \dots, Z_n(\tau) \le s_n) = \frac{\partial}{\partial\tau}\int_K f(x, \tau)dx \overset{\text{Lemma 6}}{=} \int_K \frac{\partial}{\partial\tau} f(x, \tau)dx, \tag{5}$$

where the last equation follows by Lemma 6 applied to $\lambda_1 = \cdots = \lambda_n = 0$. Therefore, it is equivalent to show that $\int_K \frac{\partial}{\partial\tau} f(x, \tau)dx \ge 0$.

Step II: We want to compute the explicit formula of $\frac{\partial}{\partial\tau} f(x, \tau)$ by identifying its Laplace transform.

Firstly, we compute the Laplace transform of $\frac{\partial}{\partial\tau} f(x, \tau)$. By Lemma 6, applied to $K = [0, \infty)^n$, we have for any $\lambda_1, \dots, \lambda_n \ge 0$,

$$\int_{[0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial}{\partial\tau} f(x, \tau)dx = \frac{\partial}{\partial\tau}\int_{[0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} f(x, \tau)dx. \tag{6}$$

However by Lemma 4, we have

$$\int_{[0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} f(x, \tau)dx = \mathbb{E}exp\Big(-\frac{1}{2}\sum_{k=1}^n \lambda_k X_k^2(\tau)\Big) = |I + \Lambda C(\tau)|^{-\frac{1}{2}}, \tag{7}$$

where Λ = diag$(\lambda_1, \dots, \lambda_n)$. Then by formula in Lemma 5 (i), we obtain

$$|I + \Lambda C(\tau)| = 1 + \sum_{\emptyset \ne J \subset [n]} |(\Lambda C(\tau))_J| = 1 + \sum_{\emptyset \ne J \subset [n]} |C(\tau)_J| \prod_{j \in J} \lambda_j. \tag{8}$$

Fix $\emptyset \ne J \subset [n] := \{1, \dots, n\}$, then $J = J_1 \cup J_2$, where $J_1 := [n_1] \cap J$, $J_2 := J - [n_1]$ and $C(\tau)_J = \begin{pmatrix} C_{J_1} & \tau C_{J_1J_2} \\ \tau C_{J_2J_1} & C_{J_2} \end{pmatrix}$. If one of J_i is $\emptyset$, then $C(\tau)_J = C_J$; otherwise, by Lemma 5.(ii) we get

$$|C(\tau)_J| = |C_{J_1}||C_{J_2}|\Big|I_{|J_1|} - \tau^2 C_{J_1}^{-\frac{1}{2}} C_{J_1J_2} C_{J_2}^{-1} C_{J_2J_1} C_{J_1}^{-\frac{1}{2}}\Big| = |C_{J_1}||C_{J_2}|\prod_{i=1}^{|J_1|}\big(1 - \tau^2\mu_{J_1,J_2}(i)\big), \tag{9}$$

where $\mu_{J_1,J_2}(i)$ denote the eigenvalues of $C_{J_1}^{-\frac{1}{2}} C_{J_1J_2} C_{J_2}^{-1} C_{J_2J_1} C_{J_1}^{-\frac{1}{2}}$. Notice that these eigenvalues belong to [0,1] by Lemma 5 (ii). Therefore, for any $\emptyset \ne J \subset [n]$ and $\tau \in [0, 1]$, we have

$$a_J(\tau) := -\frac{\partial}{\partial\tau}|C(\tau)_J| \ge 0.$$

Therefore, we can write

$$\frac{\partial}{\partial\tau}|I+\Lambda C(\tau)|^{-\frac{1}{2}} = -\frac{1}{2}|I+\Lambda C(\tau)|^{-\frac{3}{2}}\sum_{\emptyset\neq J\subset[n]}\frac{\partial}{\partial\tau}|C(\tau)_J||\Lambda_J| = \frac{1}{2}|I+\Lambda C(\tau)|^{-\frac{3}{2}}\sum_{\emptyset\neq J\subset[n]}a_J(\tau)\prod_{j\in J}\lambda_j. \tag{10}$$

We have thus shown that

$$\int_{[0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i}\frac{\partial}{\partial\tau}f(x,\tau)\,dx = \sum_{\emptyset\neq J\subset[n]}\frac{1}{2}a_J(\tau)|I+\Lambda C(\tau)|^{-\frac{3}{2}}\prod_{j\in J}\lambda_j. \tag{11}$$

Let $h_\tau := h_{3,C(\tau)}$ be the density function on $(0,\infty)^n$ defined in Lemma 8. Then by Lemmas 8 and 7 (iii), we identify the Laplace transform

$$|I+\Lambda C(\tau)|^{-\frac{3}{2}}\prod_{j\in J}\lambda_j = \int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i}\frac{\partial^{|J|}}{\partial x_J}h_\tau. \tag{12}$$

This shows that

$$\frac{\partial}{\partial\tau}f(x,\tau) = \sum_{\emptyset\neq J\subset[n]}\frac{1}{2}a_J(\tau)\frac{\partial^{|J|}}{\partial x_J}h_\tau(x). \tag{13}$$

Step III: Now we can prove $\int_K \frac{\partial}{\partial\tau}f(x,\tau)dx \geq 0$ using the explicit formula for $\frac{\partial}{\partial\tau}f(x,\tau)$ that we get above.

Recall that $a_J(\tau) \geq 0$ and observe that by Lemma 7 (ii),

$$\lim_{x_i\to 0^+}\frac{\partial^{|J|}}{\partial x_J}h_\tau(x) = 0 \quad \text{for} \quad i \notin J \subset [n].$$

Thus

$$\int_K \frac{\partial^{|J|}}{\partial x_J}h_\tau(x)\,dx = \int_{\prod_{j\in J^c}[0,t_j]} h_\tau(t_J, x_{J^c})\,dx_{J^c} \geq 0, \tag{14}$$

where $J^c = [n] - J$, and $y = (t_J, x_{J^c})$ if $y_i = t_i$ for $i \in J$ and $y_i = x_i$ for $i \in J^c$. This finishes the proof.

2.3 Auxiliary Lemmas

We shall now prove the lemmas needed for the main proof.

Lemma 4. *X is an n dimensional centered Gaussian vector with the covariance matrix C. Then for any $\lambda_1, \ldots, \lambda_n \geq 0$ we have*

$$\mathbb{E}\exp\left(-\sum_{i=1}^n \lambda_i X_i^2\right) = |I_n + 2\Lambda C|^{-1/2} \tag{15}$$

where $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Proof. We know $X \sim \mathcal{N}(0, C)$. As C is symmetric we can take A such that $AA^T = C$. Write $X \sim AY$ where $Y \sim \mathcal{N}(0, I_n)$.

Then

$$\begin{aligned}\mathbb{E}\exp\left(-\sum_{i=1}^{n}\lambda_i X_i^2\right) =&\mathbb{E}\exp(-\langle \Lambda X, X\rangle)\\ =&\mathbb{E}\exp(-\langle \Lambda AY, AY\rangle) = \mathbb{E}\exp(-\langle A^T\Lambda AY, Y\rangle)\\ =&\left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{\mathbb{R}^n} \exp\left(-\left\langle\left(\frac{1}{2}I_n + A^T\Lambda A\right)x, x\right\rangle\right)dx.\end{aligned}$$

Since $\frac{1}{2}I_n + A^T\Lambda A$ is symmetric we can write that it is equal to some UDU^T where $U \in O(n)$ and $D = \operatorname{diag}(d_1, \dots, d_n)$. Then

$$\begin{aligned}\mathbb{E}\exp\left(-\sum_{i=1}^{n}\lambda_i X_i^2\right) =&\left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{\mathbb{R}^n} \exp(-\langle Dx, x\rangle)dx\\ =&\left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \exp(-\langle Dx, x\rangle)d^n x\\ =&\prod_{i=1}^{n}\sqrt{\frac{\pi}{d_i}} = \sqrt{\frac{\pi^n}{(2\pi)^n \det D}}\\ =&2^{-n/2}\left|D\right|^{-1/2}\\ =&2^{-n/2}\left|\frac{1}{2}I_n + A^T\Lambda A\right|^{-1/2} = \left|I_n + 2A^T\Lambda A\right|^{-1/2}.\end{aligned}$$

From Sylvester's identity,

$$\left|I_n + 2A^T\Lambda A\right|^{-1/2} = \left|I_n + 2\Lambda AA^T\right|^{-1/2} = |I_n + 2\Lambda C|^{-1/2}. \qquad \square$$

Lemma 5. *For any matrix* $A \in M_{n\times n}$*,*

i)

$$|I_n + A| = 1 + \sum_{\emptyset \neq J \subset [n]} |A_J|. \tag{16}$$

ii) If $n = n_1 + n_2$*, taking the block representation of* A*,* $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix}$*, with* $A_{ij} \in M_{n_i \times n_j}$ *and* A_{11}, A_{22} *invertible, then*

$$|A| = |A_{11}||A_{22}|\left|I_{n_1} - A_{11}^{-1/2}A_{12}A_{22}^{-1}A_{21}A_{11}^{-1/2}\right|. \tag{17}$$

iii) If A *is symmetric and positively defined, then*

$$0 \leq A_{11}^{-1/2}A_{12}A_{22}^{-1}A_{21}A_{11}^{-1/2} \leq I_{n_1} \tag{18}$$

Proof.

i) Use Leibniz's formula, we get

$$\begin{aligned}|I+A| &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n}(I+A)_{\sigma_i,i}\\ &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{\sigma \text{ fixes } i}(a_{i,i}+1)\prod_{\sigma \text{ does not fix } i}(a_{\sigma_i,i})\\ &=1+\sum_{\emptyset\neq J\subset[n]}\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{\sigma \text{ does not fix } J}(a_{\sigma_i,i})\\ &=1+\sum_{\emptyset\neq J\subset[n]}|A_J|.\end{aligned}$$

ii) Note that A has decomposition

$$\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}=\begin{pmatrix}A_{11}^{1/2} & 0\\ 0 & A_{22}^{1/2}\end{pmatrix}\begin{pmatrix}I_{n_1} & A_{11}^{-1/2}A_{12}A_{22}^{-1/2}\\ A_{22}^{-1/2}A_{21}A_{11}^{-1/2} & I_{n_2}\end{pmatrix}\begin{pmatrix}A_{11}^{1/2} & 0\\ 0 & A_{22}^{1/2}\end{pmatrix},$$

then we have

$$\begin{aligned}|A| =&|A_{11}|\left|\begin{pmatrix}I_{n_1} & A_{11}^{-1/2}A_{12}A_{22}^{-1/2}\\ A_{22}^{-1/2}A_{21}A_{11}^{-1/2} & I_{n_2}\end{pmatrix}\right||A_{22}|\\ =&|A_{11}||A_{22}|\left|I_{n_1}-A_{22}^{-1/2}A_{21}A_{11}^{-1}A_{12}A_{22}^{-1/2}\right|.\end{aligned}$$

iii) Note that

$$A_{11}^{-1/2}A_{12}A_{22}^{-1}A_{21}A_{11}^{-1/2}=(A_{22}^{-1/2}A_{21}A_{11}^{-1/2})^T(A_{22}^{-1/2}A_{21}A_{11}^{-1/2})$$

is a symmetric matrix. Denote $A_{22}^{-1/2}A_{21}A_{11}^{-1/2}$ by B.

We have $x^TB^TBx=(Bx)^T(Bx)=|Bx|\geq 0$, so $B^TB\geq 0$.

For any $x,y\in\mathbb{R}^{n_1},\mathbb{R}^{n_2}$ respectively we can consider the vector $\begin{pmatrix}tx\\ y\end{pmatrix}$ where $t\in\mathbb{R}$, then we have the product

$$\begin{pmatrix}tx^T & y^T\end{pmatrix}A\begin{pmatrix}tx\\ y\end{pmatrix}=t^2x^TA_{11}x+2ty^TA_{21}x+y^TA_{22}y\geq 0.$$

This as a function of t has at most one root. That is, for any x,y,

$$(2y^TA_{21}x)^2-4x^TA_{11}xy^TA_{22}y=(y^TA_{21}x)^2-x^TA_{11}xy^TA_{22}y\geq 0.$$

Taking $x=A_{11}^{-1/2}x$ and $y=A_{22}^{-1/2}y$, we have

$$|x|^2|y|^2\geq(y^TBx)^2,$$

and when $y=Bx$,

$$|x|^2\geq x^TB^TBx.$$

Which means $x^T(I-B^TB)x\geq 0$. □

Lemma 6. $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ *is nondegenerate. Take*

$$C(\tau) := \begin{pmatrix} C_{11} & \tau C_{12} \\ \tau C_{21} & C_{22} \end{pmatrix}, 0 \le \tau \le 1,$$

and $X(\tau) \sim \mathcal{N}(0, C(\tau))$*,* $Z_i(\tau) = \frac{1}{2} X_i(\tau)^2$ *for* $i = 1, \dots, n$*.* $f(x, \tau)$ *is the density of* $Z(\tau)$*. (As defined in the beginning of the main proof)*

Then for any Borel set $K \subset [0, \infty)^n$ *and any* $\lambda_1, \dots, \lambda_n \ge 0$*,*

$$\int_K e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial}{\partial \tau} f(x, \tau) dx = \frac{\partial}{\partial \tau} \int_K e^{-\sum_{i=1}^n \lambda_i x_i} f(x, \tau) dx. \tag{19}$$

Proof. Since C is nondegenerate $x^T C x \neq 0$ for any $x \in \mathbb{R}^n$. Then $x^T C x > 0$ as C is positively defined.

Taking $u \in \mathbb{R}^{n_1}$ and $0 \in \mathbb{R}^{n_2}$, we get $\begin{pmatrix} u^T & 0 \end{pmatrix} C \begin{pmatrix} u \\ 0 \end{pmatrix} = u^T C_{11} u > 0$. Similarly $v^T C_{22} v > 0$, then $\begin{pmatrix} C_{11} & 0 \\ 0 & C_{22} \end{pmatrix}$ is strictly positively defined.

Note that

$$C(\tau) = \tau C(1) + (1 - \tau) C(0).$$

Then for any $x \in \mathbb{R}^n$,

$$x^T C(\tau) x = \tau x^T C(1) x + (1 - \tau) x^T C(0) x.$$

$C(1), C(0)$ are strictly positively defined (nondegenerate), so $x^T C(\tau) x > 0$ for any x, that is, $C(\tau)$ is nondegenerate.

Then $|C(\tau)| > 0$ for all $\tau \in [0, 1]$ and so $\frac{1}{\sqrt{|C(\tau)|}}$ is smooth.

This means $X(\tau)$ has p.d.f.

$$g(x, \tau) = \frac{1}{\sqrt{|C(\tau)|(2\pi)^n}} \exp(-\frac{1}{2} \langle C(\tau)^{-1} x, x \rangle),$$

and so $Z(\tau)$ has p.d.f.

$$\begin{aligned} f(x, \tau) =& P(Z_1 \le x_1, \dots, Z_n \le x_n) \\ =& \int_{-\sqrt{2x_1}}^{\sqrt{2x_1}} \cdots \int_{-\sqrt{2x_n}}^{\sqrt{2x_n}} g(x, \tau) d^n x \\ =& \frac{1}{\sqrt{|C(\tau)|(2\pi)^n}} \prod_{k=1}^n \frac{1}{\sqrt{2x_k}} \sum_{\epsilon \in \{-1,1\}^n} \exp(-\langle C(\tau)^{-1} y, y \rangle) \\ =& \frac{1}{\sqrt{|C(\tau)|(4\pi)^n x_1 \dots x_n}} \sum_{\epsilon \in \{-1,1\}^n} \exp(-\langle C(\tau)^{-1} y, y \rangle), \end{aligned}$$

where $y_i = \epsilon_i \sqrt{x_i}$.

We have

$$\frac{\partial}{\partial \tau}\exp(-\langle C(\tau)^{-1}y, y\rangle) = -\langle C(\tau)^{-1}(C(1) - C(0))C(\tau)^{-1}y, y\rangle\exp(-\langle C(\tau)^{-1}y, y\rangle).$$

Since $C(\tau)$ is continuous on $[0, 1]$, the largest eigenvalue λ of $C(\tau)^{-1}$ reaches some minimum on $[0, 1]$. Since $x^T C(\tau)x \leq \lambda|x|^2$ as $C(\tau)$ is symmetric and positive definite, there is some $a > 0$ such that over $[0, 1]$,

$$\langle C(\tau)^{-1}y, y\rangle \geq a\langle y, y\rangle = a\sum_{i=1}^{n} |\epsilon_i^2\sqrt{x_i}^2| = a\sum_{i=1}^{n} |x_i|.$$

Similarly, there is some $b < \infty$ such that

$$\langle C(\tau)^{-1}(C(1) - C(0))C(\tau)^{-1}y, y\rangle \leq b(y, y) = b\sum_{i=1}^{n} |x_i|.$$

Then we can bound $\frac{\partial}{\partial \tau}f(x, \tau)$ by

$$\sup_{\tau\in[0,1]}\left|\frac{\partial}{\partial \tau}f(x, \tau)dx\right| \leq \frac{1}{\sqrt{(4\pi)^n x_1 \ldots x_n}}\left(1 + b\sum_{i=1}^{n} |x_i|\right)\exp\left(-a\sum_{i=1}^{n} |x_i|\right)$$

which is integrable.

Then we have a bound on $|e^{-\sum_{i=1}^{n}\lambda_i x_i}\frac{\partial}{\partial \tau}f(x, \tau)|$.

By dominated convergence theorem, we get the result

$$\int_K e^{-\sum_{i=1}^{n}\lambda_i x_i}\frac{\partial}{\partial \tau}f(x, \tau)dx = \frac{\partial}{\partial \tau}\int_K e^{-\sum_{i=1}^{n}\lambda_i x_i}f(x, \tau)dx. \tag{20}$$

□

Lemma 7. *Let $Y = (Y_1, \ldots, Y_n)$ be a random n-dimensional vector with nonnegative coordinates. Take $\mu > 0$ and for any $\alpha = (\alpha_1, \ldots, \alpha_n) \in (0, \infty)^n$, take $x > 0, y \geq 0$ and set*

$$g_{\alpha_i}(x, y) := e^{-x-y}\sum_{k=0}^{\infty}\frac{x^{k+\alpha_i-1}}{\Gamma(k + \alpha_i)}\frac{y^k}{k!}.$$

Then define h_α as

$$h_\alpha := \mathbb{E}\left[\prod_{i=1}^{n}\frac{1}{\mu}g_{\alpha_i}\left(\frac{x_i}{\mu}, Y_i\right)\right], x_i > 0.$$

We have

i) For any α, $h_\alpha \geq 0$ and $\int_{(0,\infty)^n} h_\alpha(x)dx = 1$.

ii) If $\alpha_i > 1$, then $\lim_{x_i\to 0^+} h_\alpha(x) = 0$ and

$$\frac{\partial}{\partial x_i}h_\alpha(x) = \frac{1}{\mu}(h_{\alpha-e_i}(x) - h_\alpha(x)).$$

iii) If $\alpha_i > 1$ for all i, then for any $J \subset [n]$, $\frac{\partial^{|J|}}{\partial x_J} h_\alpha(x)$ exists and is in $L_1((0,\infty)^n)$. Moreover, for $\lambda_1, \ldots, \lambda_n \geq 0$,

$$\int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial^{|J|}}{\partial x_J} h_\alpha(x) dx = \prod_{i \in J} \lambda_i \int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x) dx.$$

Proof.

i) Since $g_{\alpha_i} \geq 0$ we have $h_\alpha \geq 0$. By Fubini theorem,

$$\begin{aligned}\int_{(0,\infty)^n} h_\alpha(x) dx =& \int_{(0,\infty)^n} \mathbb{E}\left[\prod_{i=1}^n \frac{1}{\mu} g_{\alpha_i}\left(\frac{x_i}{\mu}, Y_i\right)\right] \\ =& \mathbb{E} \prod_{i=1}^n \left[\int_0^\infty \frac{1}{\mu} g_{\alpha_i}\left(\frac{x_i}{\mu}, Y_i\right) dx\right] \\ =& \mathbb{E} \prod_{i=1}^n \left[\int_0^\infty g_{\alpha_i}(u_i, Y_i) du_i\right] \\ =& 1.\end{aligned}$$

ii) Note that $\Gamma(x)$ is decreasing on $(0,2)$ and increasing on $(2,\infty)$. Then since $\Gamma(x) > \frac{1}{2}$, we have when $k \geq 1, \alpha > 0$, and

$$\frac{\Gamma(k+\alpha)}{\Gamma(k)} \geq \frac{\frac{1}{2}}{\Gamma(1)} = \frac{1}{2}.$$

That is, $\Gamma(k+\alpha) \geq \frac{1}{2}\Gamma(k) = \frac{1}{2}(k-1)!$.

This means

$$\begin{aligned}g_{\alpha_i}(x,y) \leq& e^{-x-y} \sum_{k=0}^\infty \frac{x^{k+\alpha_i-1}}{\Gamma(k+\alpha_i)} \sum_{k=0}^\infty \frac{y^k}{k!} = e^{-x} \sum_{k=0}^\infty \frac{x^{k+\alpha_i-1}}{\Gamma(k+\alpha_i)} \\ \leq& 2e^{-x} \sum_{k=0}^\infty \frac{x^{k+\alpha_i-1}}{(k-1)!} = 2e^{-x} x^{\alpha_i} \sum_{k=0}^\infty \frac{x^{k-1}}{(k-1)!} \\ =& 2e^{-x} x^{\alpha_i}(x^{-1} + e^x) = 2x^{\alpha_i - 1}(e^{-x} + x).\end{aligned}$$

That is,

$$\begin{aligned}h_\alpha(x) =& \mathbb{E}\left[\prod_{i=1}^n \frac{1}{\mu} g_{\alpha_i}\left(\frac{x_i}{\mu}, Y_i\right)\right] \\ \leq& \prod_{i=1}^n \frac{2}{\mu}\left(\left(\frac{x_i}{\mu}\right)^{\alpha_i - 1}\left(e^{-\left(\frac{x_i}{\mu}\right)} + \left(\frac{x_i}{\mu}\right)\right)\right) \\ \leq& \left(\frac{2}{\mu}\right)^n \prod_{i=1}^n \left(\frac{x_i}{\mu}\right)^{\alpha_i - 1}\left(1 + \frac{x_i}{\mu}\right).\end{aligned}$$

Since $\alpha_i > 1$, we have a $\left(\frac{x_i}{\mu}\right)^{\alpha_i - 1}$ in the product, which approaches 0 as $x_i \to 0^+$. That is, $\lim_{x_i \to 0^+} h_\alpha(x) = 0$.

Consider the derivative of g_{α_i} with respect to x,

$$\begin{aligned}\frac{\partial}{\partial x} g_{\alpha_i} =&e^{-x-y} \sum_{k=0}^{\infty} \frac{(k+\alpha_i-1)x^{k+\alpha_i-2}}{\Gamma(k+\alpha_i)} \frac{y^k}{k!} - e^{-x-y} \sum_{k=0}^{\infty} \frac{x^{k+\alpha_i-1}}{\Gamma(k+\alpha_i)} \frac{y^k}{k!} \\ =&e^{-x-y} \sum_{k=0}^{\infty} \frac{x^{k+\alpha_i-2}}{\Gamma(k+\alpha_i-1)} \frac{y^k}{k!} - e^{-x-y} \sum_{k=0}^{\infty} \frac{x^{k+\alpha_i-1}}{\Gamma(k+\alpha_i)} \frac{y^k}{k!} \\ =&g_{\alpha_i-1} - g_{\alpha_i}.\end{aligned}$$

We have an upper bound for $g_\alpha(x,y) = |g_\alpha(x,y)|$, so by dominated convergence theorem we can differentiate under the expected value sign,

$$\begin{aligned}\frac{\partial}{\partial x_i} h_\alpha(x) =&\mathbb{E}\left[\frac{\partial}{\partial x_i} \prod_{j=1}^{n} \frac{1}{\mu} g_{\alpha_j}\left(\frac{x_j}{\mu}, Y_j\right)\right] \\ =&\mathbb{E}\left[\frac{1}{\mu^2}\left(g_{\alpha_i-1}\left(\frac{x_i}{\mu}, Y_i\right) - g_{\alpha_i}\left(\frac{x_i}{\mu}, Y_i\right)\right) \prod_{j=1, j\neq i}^{n} \frac{1}{\mu} g_{\alpha_j}\left(\frac{x_j}{\mu}, Y_j\right)\right] \\ =&\frac{1}{\mu}(h_{\alpha-e_i}(x) - h_\alpha(x)).\end{aligned}$$

iii) From ii), $\frac{\partial^{|J|}}{\partial x_J} h_\alpha(x)$ is the sum of $2^{|J|}$ different $c_I h_{\alpha_I}$s, where $\alpha_I = \alpha - \sum_{i \in I} e_i$, I ranges over all subsets of J, and c is $(-1)^{|J|-|I|}$ times a power (no more than $|J|$) of $\frac{1}{\mu}$.

This is in $L_1((0,\infty)^n)$ as we have

$$\int_{(0,\infty)^n} \left|\frac{\partial^{|J|}}{\partial x_J} h_\alpha(x)\right| \leq \sum_{I \in \mathcal{P}(J)} |c| \int_{(0,\infty)^n} |h_{\alpha_I}| = \sum_{I \in \mathcal{P}(J)} |c| \int_{(0,\infty)^n} h_{\alpha_I} = \sum_{I \in \mathcal{P}(J)} |c| < \infty.$$

Note that $\frac{\partial^{|J|}}{\partial x_J} h_\alpha = \frac{\partial}{\partial x_i} \frac{\partial^{|J| \setminus i}}{\partial x_{J \setminus i}} h_\alpha$, which is the i^{th} derivative of a linear combination of h_{α_I}s with $\alpha_{I_i} > 1$. We have that the derivatives evaluate to

$$\int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial}{\partial x_j} h_\alpha(x) dx =e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x)\Big|_{x_j=0}^{\infty} + \lambda_j \int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x) dx$$

For any $j \notin J$ we have $\lim_{x_j \to 0^+} \frac{\partial^{|J|}}{\partial x_J} h_\alpha(x)$ is the limit as x_j approaches 0^+ of a linear combination of h_{α_I}s where $\alpha_{I_j} > 1$. That is, this limit approaches 0. This means $e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x)\Big|_{x_j=0}^{\infty} = 0$, and so

$$\int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial}{\partial x_j} h_\alpha(x) dx = \lambda_j \int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x) dx$$

By induction, for any $J \subset [n]$,

$$\int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial^{|J|}}{\partial x_J} h_\alpha(x) dx = \prod_{j \in J} \lambda_j \int_0^\infty e^{-\sum_{i=1}^n \lambda_i x_i} h_\alpha(x) dx \qquad \square$$

Lemma 8. *Take* C *a strictly positively defined symmetric* $n \times n$ *matrix. Then there is some* $\mu > 0$ *with* $C - \mu I_n$ *is positively defined, meaning* $C = \mu I_n + AA^T$ *for some matrix* A*. Take random variables* $(g_j^{(l)})_{j \le n, l \le k}$ *i.i.d.* $\mathcal{N}(0,1)$*, then set*

$$Y_i = \frac{1}{2\mu} \sum_{l=1}^k \sum_{j,j' \le n} g_j^{(l)} g_{j'}^{(l)} a_{i,j} a_{i,j'} = \sum_{l=1}^k \left(\sum_{j=1}^n \frac{1}{\sqrt{2\mu}} g_j^{(l)} a_{i,j} \right)^2$$

for $i = 1, \dots, n$*. Then we take* $\alpha = (\frac{k}{2}, \dots, \frac{k}{2})$ *and* $h_{k,C} := h_\alpha$*.*

For any $\lambda_1, \dots, \lambda_n \ge 0$*, taking* $\Lambda := \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ *we have*

$$\int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} h_{k,C}(x) = |I_n + \Lambda C|^{-k/2}.$$

Note that $\int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} \frac{\partial}{\partial \tau} f(x,\tau) dx$ is the Laplace transform of $f(x,\tau)$. In this case, $\int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} h_{k,C}(x)$ is the Laplace transform of the convolution $f(x,\tau) \star f(x,\tau) \star f(x,\tau)$.

Proof. Again through Fubini theorem,

$$\begin{aligned}
\int_{(0,\infty)^n} e^{-\sum_{i=1}^n \lambda_i x_i} h_{k,C}(x) dx =& \mathbb{E}\left[\prod_{i=1}^n \int_0^\infty e^{-\lambda_i x_i} \frac{1}{\mu} g_{k/2}\left(\frac{x_i}{\mu}, Y_i \right) dx_i \right] \\
=& \mathbb{E}\left[\prod_{i=1}^n e^{-Y_i} \sum_{k=0}^\infty \frac{Y_i^k}{k!} \int_0^\infty e^{-\frac{1}{\mu} x_i - \lambda_i x_i} \frac{x^{k+\alpha-1}}{\mu^{k+\alpha} \Gamma(k+\alpha)} dx_i \right] \\
=& \mathbb{E}\left[\prod_{i=1}^n e^{-Y_i} \sum_{k=0}^\infty \frac{Y_i^k}{k!(1+\mu\lambda_i)^{k+\alpha_i}} \right] \\
=& \mathbb{E}\left[\frac{e^{-Y_i \frac{\mu\lambda_i}{1+\mu\lambda_i}}}{(1+\mu\lambda_i)^\alpha} \right] \\
=& |I_n + \mu\Lambda|^{-k/2} \mathbb{E}\left[\exp\left(-Y_i \frac{\mu\lambda_i}{1+\mu\lambda_i} \right) \right]
\end{aligned}$$

Since $Y_i = \sum X_{i_k}^2$ where $X_i \sim \mathcal{N}(0, \frac{1}{2\mu} AA^T)$, from Lemma 4 the covariance matrix of Y is $2\mu\Lambda(I + \mu\Lambda)^{-1} \frac{1}{2\mu} AA^T = \Lambda(I + \mu\Lambda)^{-1} AA^T$. Then

$$\begin{aligned}
|I_n + \mu\Lambda|^{-k/2} \mathbb{E}\left[\exp\left(-Y_i \frac{\mu\lambda_i}{1+\mu\lambda_i} \right) \right] =& |I_n + \mu\Lambda|^{-k/2} |I_n + \Lambda(I+\mu\Lambda)^{-1} AA^T|^{-k/2} \\
=& |I_n + \mu\Lambda + \Lambda(I+\mu\Lambda)(I+\mu\Lambda)^{-1} AA^T|^{-k/2} \\
=& |I_n + \Lambda C|^{-k/2} \qquad \square
\end{aligned}$$

3 Confidence sets in the Multiple Linear Regression

Given a random variable $X = (X_1, \dots, X_n)$ where X_i are i.i.d., then we can consider the parameter space Θ of X given data on X. The *confidence set*, is then a subset S of Θ, satisfying some confidence level α such that

$$\alpha = \min\{\mathbb{P}(\theta \in S) : \theta \in \Theta\}. \tag{21}$$

3.1 Confidence intervals for parameters of normal distribution

Let us consider a sample of normal distribution with mean μ and variance σ^2.

$$X_1, X_2, ..., X_n \sim N(\mu, \sigma^2).$$

Using *Maximum likelihood estimation (MLE)*, we obtain the following estimates of μ and σ^2

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2. \tag{22}$$

According to the *Law of Large Numbers (LLN)*, we know these estimates converge to μ and σ^2 as $n \to \infty$. But we also want to know: how close are these estimates to actual values of the unknown parameters μ and σ^2?

We will use confidence interval to describe precisely how close $\bar{X}$ and $\bar{X^2} - (\bar{X})^2$ are to μ and σ^2. Firstly, let's define the *Joint Distribution* of $(\bar{X}\ , \bar{X^2} - (\bar{X})^2)$.

Definitions. If $X_1, X_2, ..., X_n$ are i.i.d. standard normal, then the distribution of

$$X_1{}^2 + X_2{}^2 + ... + X_n{}^2$$

is called the χ^2_n**-distribution (chi-squared distribution)** with n degrees of freedom.

Remark. This distribution only depends on the degrees of freedom n.

Theorem 3. *If* $X_1, X_2, ..., X_n$ *are* i.i.d. *standard normal, then sample mean* $\bar{X}$ *and sample variance* $\bar{X^2} - (\bar{X})^2$ *are independent, and*

$$\sqrt{n}\bar{X} \sim N(0, 1) \quad \text{and} \quad n(\bar{X^2} - (\bar{X})^2) \sim \chi^2_{n-1}. \tag{23}$$

Proof. Consider random vector Y given by a specific orthogonal transformation of X

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = VX = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

where each v_i is a row vector. Without loss of generality, we can choose $v_1 = \left(\frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}\right)$ and let the remaining rows be any vectors such that V is orthogonal matrix. (For example, we can simply choose the rows $v_2, ..., v_n$ to be orthogonal basis in the hyperplane orthogonal to vector v_1.)

Firstly, $Y_1, ..., Y_n$ are i.i.d. standard normal, then we have $Y \sim N(V\mu_X, VIV^T) = N(0, I)$ since $\mu_X = 0$ and V is orthogonal matrix, $VV_T = I$. Because of the particular choice of the first row v_1, the $r.v.Y_1$

$$Y_1 = \frac{1}{\sqrt{n}}X_1 + \cdots + \frac{1}{\sqrt{n}}X_n = \sqrt{n}\bar{X},$$

and, therefore, $\bar{X} = \frac{1}{\sqrt{n}}Y_1$.

Next, n times sample variance can be written as

$$n(\bar{X}^2 - (\bar{X})^2) = X_1^2 + \cdots + X_n^2 - Y_1^2. \tag{24}$$

Since orthogonal transformation V preserves the length of X (i.e. $|Y| = |VX| = |X|$), therefore, we get

$$n(\bar{X}^2 - (\bar{X})^2) = Y_1^2 + \cdots + Y_n^2 - Y_1^2 = Y_2^2 + \cdots + Y_n^2 \sim \chi^2_{n-1}. \tag{25}$$

Since Y_1 and $Y_2, \ldots, Y_n$ are independent, then we get the sample mean and sample variance are independent ; $\sqrt{n}\bar{X} = Y_1 \sim N(0, 1)$ and $n(\bar{X}^2 - (\bar{X})^2) \sim \chi^2_{n-1}$. □

For general normal distribution $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \alpha^2)$, we can normalize them and then have

$$Z_1 = \frac{X_1 - \mu}{\sigma}, \ldots, Z_n = \frac{X_n - \mu}{\sigma} \overset{\text{i.i.d.}}{\sim} N(0, 1)$$

are independent standard normal. Applying theorem to $Z_1, \ldots, Z_n$ gives that

$$\sqrt{n}\bar{Z} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1), \tag{26}$$

and

$$n(\bar{Z}^2 - (\bar{Z})^2) = n\big(\frac{1}{n}\sum(\frac{X_i - \mu}{\sigma})^2 - (\frac{1}{n}\sum\frac{X_i - \mu}{\sigma})^2\big) = \frac{n(\bar{X}^2 - (\bar{X})^2)}{\sigma^2} \sim \chi^2_{n-1}. \tag{27}$$

By MLE, this result can be read as $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \bar{X}^2 - (\bar{X})^2$ are independent, and

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0, 1) \quad \text{and} \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{28}$$

Now lets construct the Construct Confidence Interval of Variance σ^2. Consider a sample $X_1, \ldots, X_n$ with distribution $\mathbb{P}_{\theta_0}$ from a parametric family $\{\mathbb{P}_{\theta_0} : \theta \in \Theta\}$, and θ_0 is unknown.

Definitions. Given a confidence level parameter $\alpha \in [0, 1]$, if there exist two statistics

$$S_1 = S_1(X_1, \ldots, X_n) \quad \text{and} \quad S_2 = S_2(X_1, \ldots, X_n)$$

such that probability

$$\mathbb{P}_{\theta_0}(S_1 \leq \theta_0 \leq S_2) = \alpha \quad (\text{or} \geq \alpha),$$

then we will call $[S_1, S_2]$ a Confidence Interval for the unknown parameter θ_0 with the confidence level α.

The interpretation of this definition is that we can guarantee with confidence α that our unknown parameter lies within the interval $[S_1, S_2]$.

For $X_1, \dots, X_n$ i.i.d. $N(\mu, \alpha^2)$, let

$$A = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0, 1), \quad B = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1},$$

and A, B are independent. Then we can rewrite A and B as

$$A = Y_1 \quad \text{and} \quad B = Y_2^2 + \cdots + Y_n^2$$

for some $Y_1, \dots, Y_n$ - i.i.d. standard normal.

Let's consider p.d.f. of χ^2_{n-1} distribution and choose points c_1 and c_2 such that the area in each tail is $\frac{(1-\alpha)}{2}$ (i.e. $\mathbb{P}(c_1 \leq B \leq c_2) = \alpha$). Therefore, we can guarantee with probability α that

$$c_1 \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq c_2. \tag{29}$$

Solving this inequality for σ^2 gives the confidence interval of σ^2 with confidence level α,

$$[\frac{n\hat{\sigma}^2}{c_2}, \frac{n\hat{\sigma}^2}{c_1}]. \tag{30}$$

Now let's construct the Confidence Interval of Mean μ.

Definitions. If $Y_0, Y_1, \dots, Y_n$ are i.i.d. standard normal, then the distribution of the random variable

$$\frac{Y_0}{\sqrt{\frac{1}{n}(Y_1^2 + \cdots + Y_n^2)}}$$

is called (Student) t_n-distribution with n degrees of freedom.

Remark. This distribution only depends on the degrees of freedom n, and the distribution is symmetric.

Using A,B defined above, we can write the formula

$$\frac{A}{\sqrt{\frac{1}{(n-1)}B}} = \sqrt{n}\frac{\hat{\mu} - \mu}{\sigma} \Big/ \sqrt{\frac{1}{n-1}\frac{n\hat{\sigma}^2}{\sigma^2}} = \frac{\sqrt{n-1}}{\hat{\sigma}}(\hat{\mu} - \mu) \sim t_{n-1}. \tag{31}$$

Then we can choose constants c and $-c$ such that the area in each tail of t-distribution is $\frac{1-\alpha}{2}$, which means the interval $[-c, c]$ gives us probability α,

$$\mathbb{P}(-c \leq \frac{\sqrt{n-1}}{\hat{\sigma}}(\hat{\mu} - \mu) \leq c) = \alpha. \tag{32}$$

Solving inequality for μ, we get the confidence interval of μ with confidence level α,

$$\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n-1}} \leq \mu \leq \hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n-1}}. \tag{33}$$

3.2 Confidence Rectangle Using Gaussian Correlation Inequality

Theorem 4 (Gaussian Correlation Inequality)**.** *Let* $X = (X_1, X_2, \ldots, X_k)$ *be the vector of random variables having the k-dimensional normal distribution with zero means, variances* $\sigma_1^2, \ldots, \sigma_k^2$*, and correlation matrix* $R = \{\rho_{ij}\}$*, then for any* $c_i > 0$*,*

$$\mathbb{P}(|X_1| \le c_1, \ldots, |X_k| \le c_k) \ge \mathbb{P}(|X_1| \le c_1)\mathbb{P}(|X_2| \le c_2, \ldots, |X_k| \le c_k). \tag{34}$$

We want to find a confidence rectangle of this sample. Firstly, we need to extend this inequality a little. It's easy to get next inequality by induction

$$\mathbb{P}(|X_1| \le c_1, \ldots, |X_k| \le c_k) \ge \prod_i \mathbb{P}(|X_i| \le c_i). \tag{35}$$

And, we also claim the following theorem.

Theorem 5. *For a positive random variable* s*, which is independent of* $X_1, X_2, \ldots, X_k$*, we have*

$$\mathbb{P}\Big(\frac{|X_1|}{s} \le c_1, \ldots, \frac{|X_k|}{s} \le c_k\Big) \ge \prod_i \mathbb{P}\Big(\frac{|X_i|}{s} \le c_i\Big). \tag{36}$$

Proof. If s is just a positive constant, then this inequality is definitely true since we can set $\{c_i' = sc_i\}$ and then apply Gaussian correlation inequality to $\{c_i\}$.

Now we prove theorem for general random variable s. By Gaussian correlation inequality, we obtain for conditional probabilities

$$\mathbb{P}\Big(|X_1| \le c_1 s, \ldots, |X_k| \le c_k s\Big|s) \ge \prod_i \mathbb{P}\Big(|X_i| \le c_i s\Big|s). \tag{37}$$

Then we can take expectation w.r.t s on both sides. Define $f_i(s) = \mathbb{P}(|z_i| \le c_i s|s)$. Then we have

$$\int f_i(s) d\mathbb{P}(s) = \mathbb{E}(P(|z_i| \le c_i s|s)) = \mathbb{P}(|z_i| \le c_i s) \tag{38}$$

Based on this notation, we can rewrite (36) as

$$\int \prod_i f_i(s) d\mathbb{P}(s) \ge \prod_i \int f_i(s) d\mathbb{P}(s). \tag{39}$$

Also notice that f_i is an increasing function in s. Therefore,

$$\int \Big(f_i(s) - f_i(t)\Big)\Big(f_j(s) - f_j(t)\Big) d\mathbb{P}(s) d\mathbb{P}(t) \ge 0, \tag{40}$$

because f_i, f_j are increasing functions and then $\Big(f_i(s) - f_i(t)\Big)$, $\Big(f_j(s) - f_j(t)\Big)$ have the same sign. We can also rewrite the above inequality as

$$\int \Big(f_i(s) - f_i(t)\Big)\Big(f_j(s) - f_j(t)\Big) d\mathbb{P}(s) d\mathbb{P}(t)$$

$$
\begin{aligned}
=&\int f_i(s)f_j(s)d\mathbb{P}(s)d\mathbb{P}(t)+\int f_i(t)f_j(t)d\mathbb{P}(t)d\mathbb{P}(s)\\
&-\int f_i(s)f_j(t)d\mathbb{P}(s)d\mathbb{P}(t)-\int f_i(t)f_j(s)d\mathbb{P}(s)d\mathbb{P}(t)\\
=&\int f_i(s)f_j(s)d\mathbb{P}(s)+\int f_i(t)f_j(t)d\mathbb{P}(t)\\
&-\int f_i(s)d\mathbb{P}(s)\int f_j(t)d\mathbb{P}(t)-\int f_i(t)d\mathbb{P}(t)\int f_j(s)d\mathbb{P}(s)\\
=&2\int f_i(s)f_j(s)d\mathbb{P}(s)-2\int f_i(s)d\mathbb{P}(s)\int f_j(s)d\mathbb{P}(s)\\
\geq&0,
\end{aligned}
$$

by Fubini's theorem and the fact that $\int d\mathbb{P}(s)=1$. This implies

$$\int f_i(s)f_j(s)d\mathbb{P}(s)\geq\int f_i(s)d\mathbb{P}(s)\int f_j(s)d\mathbb{P}(s). \tag{41}$$

Then by induction, we can easily get (39), which finishes the proof. □

Now we can use these inequalities to find confidence rectangle of unknown parameter of $Y_\nu=(X_{1\nu},X_{2\nu},\dots,X_{k\nu})$, $\nu=1,\dots,n$.

Consider a random sample of n vector $Y_\nu=(X_{1\nu},X_{2\nu},\dots,X_{k\nu})$, $\nu=1,\dots,n$, where each Y_ν has same normal distribution with unknown mean values $\mu_1,\dots,\mu_k$ and unknown variances $\sigma_1^2=\cdots=\sigma_k^2=\sigma^2$.
We can estimate σ^2 by sample variance

$$s_g^2=\frac{1}{n-1}\sum_{\nu=1}^{n}(Y_{g\nu}-\bar{Y}_g)^2, \tag{42}$$

where g is some fixed index chosen from $1,\dots$ k. Then the variables $Z_j=\sqrt{n}(\bar{Y}_j-\mu_j)$ and $s=s_g$ satisfy the assumption of Theorem 5. To find a confidence rectangle for $\mu_1,\dots,\mu_k$ with confidence level α, we can determine $c_1,\dots,c_k$ such that the right-hand side of (II) equals α. Therefore, the confidence rectangle for mean value μ is

$$\bar{Y}_i-c_is_g\sqrt{n}\leq\mu_i\leq\bar{Y}_i+c_is_g\sqrt{n},\quad i=1,\dots,k. \tag{43}$$

3.3 Multiple Linear Regression

3.3.1 Introduction

We can also generalize the idea of confidence intervals to the case where there are different parameters for each random variable.

3.3.2 Maximal likelihood estimators of parameters in multiple linear regression

Consider the model

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{44}$$

Here, X_{ij} are random variables, β_j are constants, and ϵ_i are random noise variables i.i.d. $N(0, \sigma^2)$. We shall denote the matrices Y, X, β, and ϵ and assume that X has rank p.

Then given X and Y, we can obtain an estimation for the parameters β and σ^2,

Lemma. *The MLE of* β *and* σ^2 *are*

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{2}|Y - X\hat{\beta}|^2. \tag{45}$$

Proof. The p.d.f. of Y_i is shifted from that of ϵ_i,

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - X_{i.}\beta)^2}{2\sigma^2}}.$$

Then the likelihood function of Y is

$$\mathcal{L}(\beta, \sigma^2) = \prod_{i=1}^{n} f_i(Y_i) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(Y_i - X_{i.}\beta)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{|Y - X\beta|^2}{2\sigma^2}}.$$

So maximizing the likelihood function with respect to β means minimizing $|Y - X\beta|^2$. Note that

$$\begin{aligned} |Y - X\beta|^2 =& \langle Y - \sum_{j=1}^{p} X_{\cdot j}\beta_j, Y - \sum_{j=1}^{p} X_{\cdot j}\beta_j \rangle \\ =& \langle Y, Y \rangle - 2\sum_{j=1}^{p} \beta_j \langle Y, X_{\cdot j} \rangle + \sum_{j,k=1}^{p} \beta_j \beta_k \langle X_{\cdot j}, X_{\cdot k} \rangle. \end{aligned}$$

This means when the derivative with respect to β_j is zero, we have

$$-2\langle Y, X_{\cdot j} \rangle + 2\sum_{k=1}^{p} \beta_k \langle X_{\cdot j}, X_{\cdot k} \rangle = 0$$

for each $j = 1, \dots, p$. That is,

$$X^T Y = X^T X \beta.$$

Then we get a maximal likelihood estimation for β,

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{46}$$

We can now use the log-likelihood function to get a MLE for σ,

$$\begin{aligned} \ln \mathcal{L}(\beta, \sigma^2) = & \quad -\tfrac{n}{2}\log 2\pi - \tfrac{n}{2}\log \sigma^2 - \tfrac{|Y - X\beta|^2}{2\sigma^2} \qquad (47) \\ \hat{\sigma} = & \quad \tfrac{1}{n}|Y - X\hat{\beta}|^2. \quad \square \qquad (48) \end{aligned}$$

3.3.3 Distributions and independence of MLE for parameters in multiple linear regression

We can then obtain a distribution for the parameters in the model.

Theorem 6.

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1}) \quad \textit{and} \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \tag{49}$$

and the estimates are independent.

Proof. We have $Y = X\beta + \epsilon$, that is,

$$\hat{\beta} = (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon = \beta + (X^TX)^{-1}X^T\epsilon.$$

We can then compute the mean and variance of β,

$$\begin{aligned}\mathbb{E}(\hat{\beta}) =& \mathbb{E}(\beta + (X^TX)^{-1}X^T\epsilon)\\ =& \mathbb{E}(\beta) + (X^TX)^{-1}X^T\mathbb{E}(\epsilon)\\ =& \beta.\end{aligned}$$

As X^TX is symmetric, the covariance of $\hat{\beta}$ is

$$\begin{aligned}\mathbb{E}((\hat{\beta}-\beta)(\hat{\beta}-\beta)^T) =& \mathbb{E}(((X^TX)^{-1}X^T\epsilon)((X^TX)^{-1}X^T\epsilon)^T)\\ =& (X^TX)^{-1}X^T\mathbb{E}(\epsilon\epsilon^T)X(X^TX)^{-1}\\ =& (X^TX)^{-1}X^T\sigma^2X(X^TX)^{-1}\\ =& (X^TX)^{-1}\sigma^2.\end{aligned}$$

We also have

$$Y - X\hat{\beta} = X\beta + \epsilon - (X\beta - X(X^TX)^{-1}X^T\epsilon) = (I - X(X^TX)^{-1}X^T)\epsilon.$$

Use Gram-Schmidt orthogonaliztion to write X as X_0R where $X_0 \in M_{n\times p}(\mathbb{R})$ has orthogonal columns that form an orthogonal basis, while $R \in M_{p\times p}(\mathbb{R})$ is invertible and upper triangular.

Then we have

$$\hat{\beta} - \beta = (X^TX)^{-1}X^T\epsilon = ((X_0R)^T(X_0R))^{-1}(X_0R)^T\epsilon = R^{-1}X_0^T\epsilon, \tag{50}$$

and

$$\begin{aligned}\hat{\sigma}^2 =& \tfrac{1}{n}|Y - X\hat{\beta}|^2 = \tfrac{1}{n}|(I - X(X^TX)^{-1}X^T)\epsilon|^2 & (51)\\ =& \tfrac{1}{n}|(I - (X_0R)((X_0R)^T(X_0R))^{-1}(X_0R)^T)\epsilon|^2 & (52)\\ =& \tfrac{1}{n}|(I - X_0X_0^T)\epsilon|^2. & (53)\end{aligned}$$

We can add basis elements to X_0^T to form an orthogonal matrix $A \in M_{n\times n}(\mathbb{R})$ containing an orthonormal basis.

Take $g = A\epsilon$, then $g \sim N(A\mathbf{0}, A\sigma^2A^T) = N(0, \sigma^2)$.

Indeed, each g_i are independent with one another as A is orthogonal and ϵ_is are independent.

Consider the first p elements of g, $\hat{g} = \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix}$. We have $\hat{g} = X_0^T \epsilon$ which means

$$\hat{\beta} - \beta = R^{-1}\hat{g}.$$

On the other hand, we can write $\epsilon = X_0X_0^T\epsilon + (I - X_0X_0^T)\epsilon$, and so

$$\begin{aligned} |g|^2 =& \epsilon^T\epsilon = (X_0X_0^T\epsilon + (I - X_0X_0^T)\epsilon)^T(X_0X_0^T\epsilon + (I - X_0X_0^T)\epsilon) \\ =& \epsilon^T X_0X_0^T\epsilon + \epsilon^T((I - X_0X_0^T))(I - X_0X_0^T)\epsilon \\ =& |\hat{g}|^2 + |(I - X_0X_0^T)\epsilon|^2. \end{aligned}$$

That is,

$$|(I - X_0X_0^T\epsilon|^2 = |g|^2 - |\hat{g}|^2 = g_{p+1}^2 + \cdots + g_n^2. \tag{54}$$

That is, $\hat{\beta}$ and $\hat{\sigma}$ are independent, and $\frac{n\hat{\sigma}^2}{\sigma^2} = h_{p+1}^2 + \cdots + h_n^2$ where $h_j \sim N(0, 1)$. So $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$. □

3.3.4 t-tests for linear combination of β

To get the confidence intervals of β, we shall take the linear combination of $c = (c_1, \ldots, c_p)^T$ and β,

$$c_1\beta_1 + \cdots + c_p\beta_p = c^T\beta.$$

Since $\hat{\beta}$ follows a normal distribution $c^T\hat{\beta}$ does as well. Specifically,

$$\mathbb{E}(c^T\hat{\beta}) = c^T\mathbb{E}(\hat{\beta}) = c^T\beta$$

and

$$\mathbb{E}(c^T(\hat{\beta} - \beta)(\hat{\beta} - \beta)^Tc) = c^T\mathbb{E}((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T)c = \sigma^2c^T(X^TX)^{-1}c.$$

We can take

$$\frac{c^T(\hat{\beta} - \beta)}{\sigma^2c^T(X^TX)^{-1}c} \sim N(0, 1)$$

and so we can get a t-distribution by taking

$$\left(\frac{c^T(\hat{\beta} - \beta)}{\sigma^2c^T(X^TX)^{-1}c}\right) \Bigg/ \sqrt{\frac{n\hat{\sigma}^2}{\sigma^2(n-p)}} \sim t_{n-p}. \tag{55}$$

When $c_j = 0$ for all $j \neq i$, we have

$$\left(\frac{\hat{\beta}_i - \beta_i}{(X^TX)_{ii}^{-1}}\right) \Bigg/ \sqrt{\frac{n\hat{\sigma}^2}{(n-p)}} \sim t_{n-p} \tag{56}$$

which means the α% confidence interval of β_i is

$$\hat{\beta}_i \pm t_{1-\alpha/2,n-p}\sqrt{\frac{n\hat{\sigma}^2}{(n-p)}}(X^TX)_{ii}^{-1}. \tag{57}$$

3.3.5 Joint confidence set for β and F-test

Recall that

$$|\hat{g}|^2 = \hat{g}^T\hat{g} = (\hat{\beta}-\beta)^T R^T R(\hat{\beta}-\beta) = (\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta).$$

Since $\frac{|\hat{g}|^2}{\sigma^2} = \frac{g_1^2}{\sigma^2} + \cdots + \frac{g_p^2}{\sigma^2} \sim \chi_p^2$, we have

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{\sigma^2} \sim \chi_p^2. \tag{58}$$

This is independent from $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$, so we can take

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{\sigma^2} \Bigg/ \frac{n\hat{\sigma}^2}{(n-p)\sigma^2} = \frac{(n-p)(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{np\hat{\sigma}^2} \sim F_{p,n-p}. \tag{59}$$

There is some c_α such that $F_{p,n-p}(0, c_\alpha) = \alpha$, then we have a confidence set for all β,

$$\frac{(n-p)(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{np\hat{\sigma}^2} \leq c_\alpha. \tag{60}$$

Given a null hypothesis $H_0 : \beta = \beta_0$ and alternative hypothesis $H_1 : \beta \neq \beta_0$, the $(1-\alpha)\%$ significance threshold would be c_α under the F-test. The decision rule would be

$$\begin{cases} H_0 : & \frac{(n-p)(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{np\hat{\sigma}^2} \geq c_\alpha \\ H_1 : & \frac{(n-p)(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{np\hat{\sigma}^2} \leq c_\alpha \end{cases}, \tag{61}$$

where $F_{p,n-p}(0, c_\alpha) = 1-\alpha$.

3.3.6 Simultaneous confidence set and F-test for subsets of β

Given $J = \{i_1, \ldots, i_k\} \subset [p]$, we can consider the F-test for $\beta_J = (\beta_{i_1}, \ldots, \beta_{i_k})^T$.

We can take the submatrix $(X^T X)_J^{-1}$ of $X^T X$, which is the submatrix with rows and column indices in J. Then

$$\hat{\beta}_J \sim N(\beta_J, \sigma^2 (X^T X)_J^{-1}). \tag{62}$$

Take $A_J = ((X^T X)_J^{-1})^{1/2}$, that is, $(X^T X)_J^{-1} = A_J A_J^T$. We would have

$$\hat{\beta}_J - \beta_J = A_J g_J$$

for some $g_J = (g_{i_1}, \ldots, g_{i_k})^T$, where $g_{i_j} \sim N(0, \sigma^2)$.

Similar to the section above, we can get that

$$\frac{g_{i_1}^2 + \cdots + g_{i_k}^2}{\sigma^2} \sim \chi_k^2$$

and

$$\frac{(n-p)(\hat{\beta}_J-\beta)^T (X^T X)_J^{-1}(\hat{\beta}_J-\beta)}{nk\hat{\sigma}^2} \sim F_{k,n-p} \tag{63}$$

which we can then use to construct confidence intervals and F-tests for β_J.

References

[1] Thomas Royen. A simple proof of the Gaussian correlation conjecture extended to multivariate gamma distributions. *Far East J. Theor. Stat.*, 48:139–145, 2014.

[2] Dmitry Panchenko. Gaussian correlation inequality. Unpublished.

[3] Rafał Latała and Dariusz Matlak. Royen's proof of the Gaussian correlation inequality, 2015.

[4] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[5] Dmitry Panchenko. Confidence intervals for parameters of normal distribution. MIT Open CourseWare, 2006.

[6] Dmitry Panchenko. Multiple linear regression. MIT Open CourseWare, 2006.