Multivariate Statistics

Lecture 06

Fudan University

Outline

Outline

Efficiency

If a p-component random vector $\mathbf{y}$ has mean vector $\mathbb{E}[\mathbf{y}] = \boldsymbol{\nu}$ and covariance matrix $\mathbb{E}\left[(\mathbf{y}-\boldsymbol{\nu})(\mathbf{y}-\boldsymbol{\nu})^\top\right] = \boldsymbol{\Psi} \succ \mathbf{0}$, then

$$\left\{\mathbf{z} : (\mathbf{z}-\boldsymbol{\nu})^\top \boldsymbol{\Psi}^{-1}(\mathbf{z}-\boldsymbol{\nu}) = p+2\right\}$$

is called the concentration ellipsoid of $\mathbf{y}$.

Let $\boldsymbol{\theta}$ be a vector of p parameters in a distribution, and let $\mathbf{t}$ be a vector of unbiased estimators (that is, $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$) based on N observations from that distribution with covariance matrix $\boldsymbol{\Psi}$. Then the ellipsoid

$$\left\{\mathbf{z} : N(\mathbf{z}-\boldsymbol{\theta})^\top \mathbb{E}\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right](\mathbf{z}-\boldsymbol{\theta}) = p+2\right\}$$

lies entirely within the ellipsoid of concentration of $\mathbf{t}$, where f is the density of the distribution (or probability function) with respect to the components of $\boldsymbol{\theta}$.

Efficiency

The ellipsoid

$$\left\{ \mathbf{z} : N(\mathbf{z}-\boldsymbol{\theta})^\top \mathbb{E}\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] (\mathbf{z}-\boldsymbol{\theta}) = p+2 \right\}$$

lies entirely within the ellipsoid of concentration of $\mathbf{t}$

$$\left\{ \mathbf{z} : (\mathbf{z}-\boldsymbol{\theta})^\top \left(\mathbb{E}\left[(\mathbf{t}-\boldsymbol{\theta})(\mathbf{t}-\boldsymbol{\theta})^\top \right] \right)^{-1} (\mathbf{z}-\boldsymbol{\theta}) = p+2 \right\},$$

that is

$$\left(N\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top \right] \right)^{-1} \preceq \mathbb{E}\left[(\mathbf{t}-\boldsymbol{\theta})(\mathbf{t}-\boldsymbol{\theta})^\top \right].$$

Efficiency

Let $\boldsymbol{\theta}$ be a vector of p parameters in a distribution, and let $\mathbf{t}$ be a vector of unbiased estimators (that is, $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$) based on N observations from that distribution with covariance matrix $\boldsymbol{\Psi}$. Then the ellipsoid

$$\left\{\mathbf{z} : N(\mathbf{z}-\boldsymbol{\theta})^\top \mathbb{E}\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right](\mathbf{z}-\boldsymbol{\theta}) = p+2\right\} \quad (1)$$

lies entirely within the ellipsoid of concentration of $\mathbf{t}$, where f is the density of the distribution (or probability function) with respect to the components of $\boldsymbol{\theta}$.

1. If the ellipsoid (1) is the ellipsoid of concentration of $\mathbf{t}$, then $\mathbf{t}$ is said to be efficient.
2. In general, the ratio of the volume of (1) to that of the ellipsoid of concentration defines the efficiency of $\mathbf{t}$.

Efficiency

Consider the case of the multivariate normal distribution.

1. If $\boldsymbol{\theta} = \boldsymbol{\mu}$, then $\bar{\mathbf{x}}$ is efficient.
2. If $\boldsymbol{\theta}$ includes both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then $\bar{\mathbf{x}}$ and $\mathbf{S}$ have efficiency $((N-1)/N)^{p(p+1)/2}$.
3. If the normal distribution is non-singular, we have

$$\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\top}\right] = -\mathbb{E}\left[\frac{\partial^2 \ln f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\top}}\right].$$

Multivariate Cramer-Rao Inequality

Theorem 2

Under the regularity condition (everything is well-defined, integration and differentiation can be swapped), we have

$$N\mathbb{E}\left[(\mathbf{t}-\boldsymbol{\theta})(\mathbf{t}-\boldsymbol{\theta})^\top\right] \succeq \left(\mathbb{E}\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ln f(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right]\right)^{-1},$$

where $\mathbb{E}[\mathbf{t}] = \boldsymbol{\theta}$ and $f(\mathbf{x},\boldsymbol{\theta})$ is the density of the distribution with respect to the components of $\boldsymbol{\theta}$.

1. Let $\mathbf{s} = \frac{\partial \ln g(\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, where g is the density on N samples and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.
2. For unbiased estimator $\mathbf{t}$ of $\boldsymbol{\theta}$, we have $\mathrm{Cov}[\mathbf{t},\mathbf{s}] = \mathbf{I}$.

Outline

Consistency

A sequence of vectors $\mathbf{t}_n = [t_{1n}, \ldots, t_{pn}]^\top$ for $n = 1, 2, \ldots,$ is a consistent estimator of $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^\top$ if

$$\lim_{n\to\infty} t_{in} = \theta_i$$

for $i = 1, \ldots, p$.

1. By the law of large numbers, the sample mean $\bar{\mathbf{x}}$ is a consistent estimator of $\boldsymbol{\mu}$ if the observations are i.i.d with mean $\boldsymbol{\mu}$ (normality is not involved).
2. The sample covariance matrix is also consistent since

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^\top - \frac{N}{N-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top.$$

Outline

Asymptotic Normality

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with the same arbitrary distribution, zero mean, and variance σ^2.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then the random variable

$$Z = \lim_{n\to\infty} \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

is a standard normal distribution.

What about multivariate case?

Asymptotic Normality

Multivariate central limit theorem.

Theorem 3

Let p-component vectors $\mathbf{y}_1, \mathbf{y}_2, \dots$ be i.i.d with means $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ and covariance matrices $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$. Then the limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})$$

as $n \to +\infty$ is $\mathcal{N}(\mathbf{0}, \mathbf{T})$.

Characteristic Function and Probability

If $\mathbf{x}$ does not have a density, the characteristic function uniquely defines the probability of any continuity interval.

Theorem 5

Let $\{F_j(\mathbf{x})\}$ be a sequence of cdfs, and let $\{\phi_j(\mathbf{t})\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(\mathbf{x})$ to converge to a cdf $F(\mathbf{x})$ is that, for every $\mathbf{t}$, $\phi_j(\mathbf{t})$ converges to a limit $\phi(\mathbf{t})$ that is continuous at $\mathbf{t} = \mathbf{0}$. When this condition is satisfied, the limit $\phi(\mathbf{t})$ is identical with the characteristic function of the limiting distribution $F(\mathbf{x})$.

See the proof in Section 10.7 of "Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press"

Asymptotic Normality

Let

$$\mathbf{A}(n) = \sum_{\alpha=1}^{N} \left(\mathbf{x}_\alpha - \bar{\mathbf{x}}_N\right)\left(\mathbf{x}_\alpha - \bar{\mathbf{x}}_N\right)^\top,$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independently distributed according to $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $n = N - 1$. Then the limiting distribution of

$$\mathbf{B}(n) = \frac{1}{\sqrt{n}}\left(\mathbf{A}(n) - n\boldsymbol{\Sigma}\right)$$

is normal with mean $\mathbf{0}$ and covariance $\mathbb{E}\left[b_{ij}(n) b_{kl}(n)\right] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$.

Outline

Decision Theory

1. An observation random vector $\mathbf{x}$ whose distribution $P_{\boldsymbol{\theta}}$ depends on a parameter $\boldsymbol{\theta}$ which is an element of a set $\boldsymbol{\Theta}$.
2. The statistician is to make a decision $\mathbf{d}$ in a set $\mathcal{D}$.
3. A decision procedure is a function $\boldsymbol{\delta}(\cdot)$ whose domain is the set of values of $\mathbf{x}$ and whose range is $\mathcal{D}$.
4. The loss in making decision $\mathbf{d}$ for the distribution of $\mathbf{x}$ is a nonnegative function $L(\boldsymbol{\theta}, \mathbf{d})$.
5. The evaluation of a procedure $\boldsymbol{\delta}(\mathbf{x})$ is on the basis of the risk function

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathbb{E}_{\mathbf{x}\sim P_{\boldsymbol{\theta}}}\left[L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))\right].$$

For example, the risk can be the mean squared error for univariate case

$$R(\theta, \delta) = \mathbb{E}_{\mathbf{x}\sim P_{\boldsymbol{\theta}}}\left[(\delta(\mathbf{x}) - \theta)^2\right]$$

Decision Theory

1. A decision procedure $\boldsymbol{\delta}(\mathbf{x})$ is as good as a procedure $\boldsymbol{\delta}^*(\mathbf{x})$ if

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) \leq R(\boldsymbol{\theta}, \boldsymbol{\delta}^*),$$

   and $\boldsymbol{\delta}(\mathbf{x})$ is better than $\boldsymbol{\delta}^*(\mathbf{x})$ if it holds with a strict inequality for at least one value of $\boldsymbol{\theta}$.

2. A procedure $\boldsymbol{\delta}^*(\mathbf{x})$ is inadmissible if there exists another procedure $\boldsymbol{\delta}(\mathbf{x})$ that is better than $\boldsymbol{\delta}^*(\mathbf{x})$.
3. A procedure is admissible if it is not inadmissible (i.e., if there is no procedure better than it) in terms of the given loss function.

Bayes Procedure

If the parameter $\boldsymbol{\theta}$ can be assigned an a prior distribution, say, with density $\rho(\boldsymbol{\theta})$, then the average loss from use of a decision procedure $\boldsymbol{\delta}(\mathbf{x})$ is

$$r(\rho, \boldsymbol{\delta}) = \mathbb{E}_{\rho}\left[R(\boldsymbol{\theta}, \boldsymbol{\delta})\right] = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\Big[\mathbb{E}_{\mathbf{x}\sim P_{\boldsymbol{\theta}}}\left[L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))\right]\Big].$$

Given the a prior density ρ, the decision procedure $\boldsymbol{\delta}(\mathbf{x})$ that minimizes $r(\rho, \boldsymbol{\delta})$ is the Bayes procedure, and the resulting minimum of $r(\rho, \boldsymbol{\delta})$ is the Bayes risk.

Bayes Procedure

If the density of $\mathbf{x}$ given $\boldsymbol{\theta}$ is $f(\mathbf{x} \mid \boldsymbol{\theta})$, the joint density of $\mathbf{x}$ and $\boldsymbol{\theta}$ is $f(\mathbf{x} \mid \boldsymbol{\theta})\rho(\boldsymbol{\theta})$ and the average risk of a procedure $\boldsymbol{\delta}(\mathbf{x})$ is

$$\begin{aligned} r(\rho, \boldsymbol{\delta}) &= \int_\Theta \int_\mathcal{X} L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f(\mathbf{x} \mid \boldsymbol{\theta})\rho(\boldsymbol{\theta}) \,\mathrm{d}\mathbf{x}\,\mathrm{d}\boldsymbol{\theta} \\ &= \int_\mathcal{X} \left(\int_\Theta L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) g(\boldsymbol{\theta} \mid \mathbf{x}) \,\mathrm{d}\boldsymbol{\theta} \right) f(\mathbf{x}) \,\mathrm{d}\mathbf{x}, \end{aligned} \tag{2}$$

where

$$f(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid \boldsymbol{\theta})\rho(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \quad \text{and} \quad g(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \boldsymbol{\theta})\rho(\boldsymbol{\theta})}{f(\mathbf{x})}$$

are the marginal density of $\mathbf{x}$ and the a posterior density of $\boldsymbol{\theta}$ given $\mathbf{x}$.

The procedure that minimizes $r(\rho, \boldsymbol{\delta})$ is one that for each $\mathbf{x}$ minimizes the expression in braces on the right-hand side of (2), that is, the expectation of $L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))$ with respect to the a posterior distribution.

If $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are vectors and $L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) = (\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))$, where $\mathbf{Q}$ is positive definite. Then we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))\right] =& \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))\right] \\
=& \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])^\top \mathbf{Q}(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])\right] \\
&+ \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])^\top \mathbf{Q}(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))\right] \\
&+ \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])\right] \\
&+ \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))\right] \\
=& \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])^\top \mathbf{Q}(\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}])\right] \\
&+ \mathbb{E}_{\boldsymbol{\theta}|\mathbf{x}}\left[(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}] - \boldsymbol{\delta}(\mathbf{x}))\right]
\end{aligned}
$$

and the minimum occurs at $\boldsymbol{\delta}(\mathbf{x}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{x}]$ the mean of the a posterior distribution.

Bayes Procedure

If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independently distributed, each $\mathbf{x}_\alpha$ according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is normal with mean

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu} \tag{3}$$

and covariance matrix

$$\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

If the loss function is

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) = (\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))^\top \mathbf{Q}(\boldsymbol{\theta} - \boldsymbol{\delta}(\mathbf{x}))$$

then the Bayes estimator of $\boldsymbol{\mu}$ is (3).

Outline

The Biased Estimator

The sample mean $\bar{\mathbf{x}}$ seems the natural estimator of the population mean $\boldsymbol{\mu}$ based on a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

However, Stein (1956) showed $\bar{\mathbf{x}}$ is not admissible with respect to the mean squared loss when $p \geq 3$.

The Biased Estimator

Consider the loss function

$$L(\boldsymbol{\mu}, \mathbf{m}) = \|\boldsymbol{\mu} - \mathbf{m}\|_2^2\,,$$

where $\mathbf{m}$ is an estimator of the mean $\boldsymbol{\mu}$.

If $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independently distributed to $\mathcal{N}_p(\boldsymbol{\mu}, N\mathbf{I})$, we have

$$\mathbb{E}\left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2\right] = \sum_{\alpha=1}^{p} \mathrm{Var}(\bar{x}_\alpha) = p.$$

The Biased Estimator

The estimator proposed by James and Stein is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}$$

where $\boldsymbol{\nu}$ is an arbitrary fixed vector and $p \geq 3$.

It holds that $\mathbb{E}\left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right] < \mathbb{E}\left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2\right]$.

For small values of $\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2$, the multiplier of $(\bar{\mathbf{x}} - \boldsymbol{\nu})$ is negative; that is, the estimator $m(\bar{\mathbf{x}})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\bar{\mathbf{x}}$.

The Biased Estimator

Table 3.2 gives values of the risk for $p = 10$ and $\sigma^2 = 1$. For example, if $\tau^2 = \|\boldsymbol{\mu} - \boldsymbol{\nu}\|^2$ is 5, the mean squared error of the James–Stein estimator is 8.86, compared to 10 for the natural estimator; this is the case if $\mu_i - \nu_i = 1/\sqrt{2} = 0.707$, $i = 1, \ldots, 10$, for instance.

Table 3.2†. Average Mean Squared Error of the James–Stein Estimator for $p = 10$ and $\sigma^2 = 1$

$\tau^2 = \|\boldsymbol{\mu} - \boldsymbol{\nu}\|^2$	$\mathscr{E}_{\boldsymbol{\mu}}\|\boldsymbol{m}(Y) - \boldsymbol{\mu}\|^2$
0.0	2.00
0.5	4.78
1.0	6.21
2.0	7.51
3.0	8.24
4.0	8.62
5.0	8.86
6.0	9.03

† From Efron and Morris (1977).

The Biased Estimator

The estimator proposed by James and Stein is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

For small values of $\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2$, the multiplier of $(\bar{\mathbf{x}} - \boldsymbol{\nu})$ is negative; that is, the estimator $m(\bar{\mathbf{x}})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\bar{\mathbf{x}}$.

We can improve $\mathbf{m}(\bar{\mathbf{x}})$ by using

$$\tilde{\mathbf{m}}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)^+ (\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

which holds that $\mathbb{E}\left[\|\tilde{\mathbf{m}}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right] \leq \mathbb{E}\left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right]$.

Outline

Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent, standard normal random variables, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the (central) chi-squared distribution (χ^2-distribution) with n degrees of freedom.

We have $\mathbb{E}[y] = n$ and $\mathrm{Var}[y] = 2n$.

Chi-Squared Distribution

The probability density function of the (central) chi-squared distribution is

$$f(y;\, n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} \exp\left(-\dfrac{y}{2}\right), & y > 0; \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t)\, \mathrm{d}t.$$

Chi-Squared Distribution

The derivation for the density is based on

1. We have $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.
2. For $y_1 = x^2$ with $x \sim \mathcal{N}(0,1)$, the density function of y_1 is
$$\frac{1}{\sqrt{2\pi y_1}} \exp\left(-\frac{1}{2} y_1\right).$$
3. For beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$, we have
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$
4. If $F(z) = \int_{a(z)}^{b(z)} f(y,z)\,\mathrm{d}y$, then
$$F'(z) = \int_{a(z)}^{b(z)} \frac{\partial f(y,z)}{\partial z}\,\mathrm{d}x + f(b(z),z)b'(z) - f(a(z),z)a'(z).$$

Noncentral Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent and each x_i are normally distributed random variables with means μ_i and unit variances, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the noncentral Chi-squared distribution with n degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{n} \mu_i^2.$$

We have $\mathbb{E}[y] = n + \lambda$ and $\mathrm{Var}[y] = 2n + 4\lambda$.

Noncentral Chi-Squared Distribution

If $y_1, \dots, y_k$ are independent and each y_i is distributed according to the noncentral chi-squared distribution with n_i degrees of freedom and noncentrality parameter λ_i, then

$$\sum_{i=1}^{k} y_i \sim \chi^2_{n_1+\cdots+n_k}\left(\sum_{i=1}^{k} \lambda_i\right).$$

Theorem 4

If the n-component vector $\mathbf{y}$ is distributed according to $\mathcal{N}(\boldsymbol{\nu}, \mathbf{T})$ with $\mathbf{T} \succ \mathbf{0}$, then

$$\mathbf{y}^\top \mathbf{T}^{-1} \mathbf{y} \sim \chi^2_n\left(\boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu}\right).$$

If $\boldsymbol{\nu} = \mathbf{0}$, the distribution is the central χ^2-distribution.