多元正态分布参数的估计和数据的清洁与变换

张伟平

zwp@ustc.edu.cn

Office: 东区管理科研楼 1006

Phone: 63600565

课件 http://staff.ustc.edu.cn/~zwp/

论坛 http://fisher.stat.ustc.edu.cn

简介

多元正态分布的参数 μ 和 Σ 可以使用不同的统计推断方法来估计.

1.1 最大似然估计

设 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, 则负对数似然函数 为

$$
\begin{aligned}
l(\mu, \Sigma) &\propto \frac{n}{2} log|\Sigma| + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'\Sigma^{-1}(\mathbf{x} - \mu) \\
&= \frac{n}{2} log|\Sigma| + \frac{1}{2} tr[\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})] \\
&\quad + \frac{n}{2}(\mu - \bar{\mathbf{x}})'\Sigma^{-1}(\mu - \bar{\mathbf{x}})
\end{aligned}
$$

最大化似然函数等价于最小化上述函数. 令 $\frac{\partial l(\mu,\Sigma)}{\partial \mu} = 0$ 和 $\frac{\partial l(\mu,\Sigma)}{\partial \Sigma} = 0$, (忽略 Σ 的对称性) 我们得到

$$-\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu) = 0,$$

$$n\Sigma^{-1} - \Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)'\Sigma^{-1} = 0$$

从而得到解

$$\hat{\mu} = \bar{\mathbf{x}}, \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x} - \hat{\mu})(\mathbf{x} - \hat{\mu})'$$

注意到

$$l(\mu, \Sigma) \geq l(\hat{\mu}, \Sigma), \text{等号成立当且仅当} \quad \mu = \hat{\mu}$$

从而

$$\min_{\mu,\Sigma>0} l(\mu, \Sigma) = \min_{\Sigma>0} \quad l(\hat{\mu}, \Sigma)$$

下面的引理证明了 $\hat{\Sigma}$ 是 $(\hat{\mu}, \Sigma)$ 的最小值点. 从而得到最大似然估计

$$\hat{\mu}_{mle} = \bar{\mathbf{x}}, \qquad \hat{\Sigma}_{mle} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

其中 $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

引理 1. *设 B 为 p 阶正定矩阵, $n > 0$ 为实数, 在对所有 p 阶正定矩阵 Σ 有*

$$\frac{n}{2}log|\Sigma| + \frac{1}{2}tr[\Sigma^{-1}B] \geq \frac{n}{2}log|B| + \frac{pn}{2}(1 - logn)$$

当且仅当 $\Sigma = \frac{1}{n}B$ 时等号成立.

证明. 由 $B > 0$, 故存在可逆对称阵 C, 使得 $B = CC$. 记 $\tilde{\Sigma} = C^{-1}\Sigma C^{-1}$, 则 $|\Sigma| = |B||\tilde{\Sigma}|$, 有

$$\frac{n}{2}log|\Sigma| + \frac{1}{2}tr[\Sigma^{-1}B] = \frac{n}{2}log|B| + \frac{n}{2}log|\tilde{\Sigma}| + \frac{1}{2}tr[\tilde{\Sigma}^{-1}]$$

$$= \frac{n}{2}log|B| + \frac{1}{2}\sum_{i=1}^{p}[\frac{1}{\lambda_i} + nlog\lambda_i],$$

其中 $\lambda_1 \geq \cdots \geq \lambda_p > 0$ 为 $\tilde{\Sigma}$ 的特征根, 于是

$$\min_{\Sigma>0}\{\frac{n}{2}log|\Sigma| + \frac{1}{2}tr[\Sigma^{-1}B]\} = \frac{n}{2}log|B| + \frac{1}{2}\min_{\lambda_j>0}\sum_{i=1}^{p}[\frac{1}{\lambda_i} + nlog\lambda_i]$$

注意到函数 $g(\lambda) = \frac{1}{\lambda} + nlog(\lambda)$ 在 $\lambda = 1/n$ 处达到极小值, 故上式当 $\lambda_1 = \cdots \lambda_p = \frac{1}{n}$ 时达到极小值, 即 $\tilde{\Sigma} = \frac{1}{n}I_p$, 等价地

$$\hat{\Sigma} = C\tilde{\Sigma}C = \frac{1}{n}CC = \frac{1}{n}B.$$

□

使用上述引理来证明 Σ 的最大似然估计时候, 需要 $B = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}}) > 0$. 这是一随机矩阵, 因此我们需要证明当 $n > p$ 时,$P(B > 0) = 1$.

定理 1. 设样本 $X_1, \ldots, X_n$ $i.i.d \sim N_p(\mu, \Sigma)$, 记 $B = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, 则

(1) $\bar{\mathbf{x}}$ 与 B 相互独立, 且 $\bar{\mathbf{x}} \sim N_p(\mu, \frac{1}{n}\Sigma)$;

(2) $P(B > 0)$ 的充要条件是 $n > p$.

(3) $\bar{\mathbf{x}}$ 和 B 为充分完备统计量.

证明. (1) 记 $X = [X_1, \ldots, X_p]$ 为 $n \times p$ 阶矩阵, 则 $X \sim N_{n\times p}(\mathbf{1}_n\mu', I_n \otimes \Sigma)$. 再记 Γ 为 n 阶正交矩阵, 其最后一行为 $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$. 作变换

$$Z = \Gamma X := [z_1, \ldots, z_n]'$$

于是 $Z \sim N_{n\times p}(\Gamma\mathbf{1}_n\mu', I_n \otimes \Sigma)$, 因此 $z_1, \ldots, z_n$ 相互独立, 注意到

$$\Gamma\mathbf{1}_n\mu' = (0, \ldots, 0, \sqrt{n})'\mu' = (\mathbf{0}, \ldots, \mathbf{0}, \sqrt{n}\mu)'$$

所以 $z_i \sim N_p(\mathbf{0}, \Sigma), j = 1, \ldots, n-1$, $z_n \sim N_p(\sqrt{n}\mu, \Sigma)$. 而

$$\bar{\mathbf{x}} = \frac{1}{n}X'\mathbf{1}_n = \frac{1}{n}Z'\Gamma\mathbf{1}_n = \frac{1}{n}Z'(0, \ldots, 0, \sqrt{n})' = z_n/\sqrt{n}$$

$$
\begin{aligned}
B &= X'(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')X = Z'\Gamma(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\Gamma'Z \\
&= Z'Z - \frac{1}{n}(Z'\Gamma\mathbf{1}_n)(Z'\Gamma\mathbf{1}_n)' = \sum_{i=1}^{n-1} z_i z_i'.
\end{aligned}
$$

从而 $\bar{\mathbf{x}}$ 和 B 相互独立.

(2) 记 $(n-1)\times p$ 矩阵 $Z_* = (z_1, \ldots, z_{n-1})'$, 则 $B = Z_*'Z_*$, 且 $Rank(B) = Rank(Z_*)$, 于是命题 (2) 等价于要证明 $P(Rank(Z_*) = p) = 1 \Leftrightarrow n > p$. 必要性显然. 现证充分性. 若 $n > p$, 由于增加行不会导致 Z_* 的秩减少, 因此只需证明 $n = p+1$ 时满秩即可. 由

$$
\begin{aligned}
&P(z_1, \ldots, z_p\text{线性相关}) \\
&\le \sum_{i=1}^{p} P(z_i\text{为}z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_p\text{的线性组合}) \\
&= pP(z_1\text{为}z_2, \ldots, z_p\text{的线性组合}) = 0.
\end{aligned}
$$

最后一式由 $Z_1, Z_2, \ldots, Z_p$ 相互独立同分布可知不可能为线性相关.

(3) 由因子分解定理可证充分性, 完备性根据指数族性质可得. □

在有了 μ 和 Σ 的最大似然估计 $\hat{\mu}$ 和 $\hat{\Sigma}$ 后, 我们是否可以通过使用 $\hat{\mu}$ 和 $\hat{\Sigma}$ 替换前面定义过的回归系数, 相关系数, 条件协方差阵和偏相关系数等中的 μ 和 Σ 来得到相应的最大似然估计? 由下面引理知道这是可以的.

引理 2. *设 θ 的最大似然估计为 $\hat{\theta}$, 若 $\theta \to \phi(\theta)$ 为一一变换, 则 $\phi(\theta)$ 的最大似然估计为 $\phi(\hat{\theta})$.*

求相关系数的最大似然估计.

↑Example

↓Example

解 由相关系数 $R = D^{-1}\Sigma D^{-1}$,$\Sigma \to (D, R)$ 为一一变换, 因此由 Σ 的最大似然估计为 $\hat{\Sigma}_{mle}$ 知 D 的最大似然估计为 $\hat{D} = \sqrt{diag(\Sigma)}$, 从而 R 的最大似然估计为

$$\hat{R} = \hat{D}\hat{\Sigma}_{mle}\hat{D}.$$

其 (i,j) 元为

$$\hat{\rho}_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_{i\cdot})(x_{jk}-\bar{x}_{i\cdot})}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\bar{x}_{i\cdot})^2\sum_{k=1}^{n}(x_{jk}-\bar{x}_{i\cdot})^2}}.$$

↑Example

求回归系数 $\beta_{1\cdot 2}=\Sigma_{12}^{-1}\Sigma_{22}^{-1}$ 和条件协方差函数 $\Sigma_{11\cdot 2}$ 的最大似然估计.

↓Example

解 由于 $\Sigma\to(\Sigma_{11},\Sigma_{12},\Sigma_{21},\Sigma_{22})$ 为一一映射, 所以由 Σ 的最大似然估计为 $\hat{\Sigma}_{mle}=\hat{\Sigma}$ 知 $\beta_{1\cdot 2}$ 和 $\Sigma_{11\cdot 2}$ 的最大似然估计为

$$\hat{\beta}_{1\cdot 2,mle}=\hat{\Sigma}_{12}^{-1}\hat{\Sigma}_{22}^{-1}$$

$$\Sigma_{11\cdot 2,mle}=\hat{\Sigma}_{11}-\hat{\Sigma}_{12}\hat{\Sigma}_{22}\hat{\Sigma}_{21}$$

1.2 最大似然估计的性质

讨论估计量的性质常常考虑无偏性, 有效性, 相合性和渐近正态性等等.

定理 2. *在前面假设及记号下,*

$$E\hat{\mu}_{mle} = \mu, \qquad E\hat{\Sigma}_{mle} = \frac{n-1}{n}\Sigma.$$

证明. $\hat{\mu}_{mle}$ 的无偏性显然. 由前面的证明中知道

$$\hat{\Sigma}_{mle} = \frac{1}{n}\sum_{i=1}^{n-1} z_i z_i'$$

其中 $z_1, \ldots, z_{n-1} i.i.d \sim N(0, \Sigma)$. 因此

$$E\hat{\Sigma}_{mle} = \frac{1}{n}\sum_{i=1}^{n} Ez_i z_i' = \frac{n-1}{n}\Sigma.$$

□

由于 $\hat{\Sigma}_{mle}$ 不是 Σ 的无偏估计, 但可校正为无偏估计, 即为常用的**样本协方差矩阵**:

$$S = \frac{n}{n-1}\hat{\Sigma}_{mle} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

定理 3. *在前面假设及记号下易知 $(\bar{\mathbf{x}}, S)$ 为 (μ, Σ) 的一致最小方差无偏估计 (UMVUE).*

证明略.

定理 4. *在前面假设及记号下, $\hat{\mu}_{mle}, \hat{\Sigma}_{mle}$ 分别为 μ, Σ 的强 (弱) 相合估计.*

证明. 由于 $\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, 从而由 Kolmogorov 强大数律 (若 $\{X_i\}$ 为相互独立同分布的随机变量, 则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} c \Longleftrightarrow E|X_1| < \infty, c = EX_1$) 显然. 又

$$\hat{\Sigma}_{mle} = \frac{1}{n}B = (\frac{1}{n}B_{ij})$$

$$\begin{aligned}\frac{1}{n}B_{ij} &= \frac{1}{n}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)\\ &= \frac{1}{n}\sum_{k=1}^{n}x_{ki}x_{kj}-\left(\frac{1}{n}\sum_{k=1}^{n}x_{ki}\right)\left(\frac{1}{n}\sum_{k=1}^{n}x_{kj}\right)\end{aligned}$$

显然 $\frac{1}{n}\sum_{k=1}^{n}x_{ki}\overset{a.s.}{\to}\mu_i$, $\frac{1}{n}\sum_{k=1}^{n}x_{kj}\overset{a.s.}{\to}\mu_j$, 而

$$E(x_{ki}x_{kj})=E(x_{ki}-\mu_i)(x_{kj}-\mu_j)+\mu_i\mu_j=\sigma_{ij}+\mu_i\mu_j,$$

注意到 $E|x_{ki}x_{kj}|\le[Ex_{ki}^2Ex_{kj}^2]^{1/2}=[(\sigma_{ii}+\mu_i^2)(\sigma_{jj}+\mu_j^2)]^{1/2}<\infty$,
因此由 Kolomogorov 强大数律有

$$\frac{1}{n}B_{ij}\overset{a.s.}{\to}\sigma_{ij}.$$

因此 $\hat{\Sigma}_{mle}\overset{a.s.}{\to}\Sigma$. □

定理 5. 设 $X_1, \ldots, X_n, \ldots i.i.d \sim N_p(\mu, \Sigma)$, 记 $B_n = \sum_{i=1}^{n}(X_i - \bar{\mathbf{x}})(X_i - \bar{\mathbf{x}})'$, 则

$$\frac{1}{n}(B_n - n\Sigma) \xrightarrow{d} N(0, V)$$

这里指将 B_n 的独立元拉直为一个向量满足渐近正态性. V 见下.

证明. 由于 $B_n = \sum_{k=1}^{n-1} z_k z_k'$, 其中 $z_1, \ldots, z_{n-1} i.i.d \sim N(0, \Sigma)$. 将 $z_k z_k'$ 的下三角元按列排成 $p(p+1)/2$ 向量

$$Y_k = (z_{k1}^2, z_{k1}z_{k2}, \ldots, z_{k1}z_{kp}; \cdots, z_{k2}^2, \ldots, z_{kp}^2)',$$

则 Y_i 的矩可以得出

$$Ez_{ki}z_{kj} = \sigma_{ij}, \quad Ez_{ki}z_{kj}z_{ks}z_{kt} = \sigma_{ij}\sigma_{st} + \sigma_{is}\sigma_{jt} + \sigma_{it}\sigma_{js}$$

从而

$$cov(z_{ki}z_{kj}, z_{ks}z_{kt}) = (z_{ki}z_{kj} - \sigma_{ij})(z_{ks}z_{kt} - \sigma_{lt}) = \sigma_{is}\sigma_{jt} + \sigma_{it}\sigma_{js}.$$

因此由中心极限定理易知

$$\frac{1}{n}(B_n - n\Sigma) = \frac{n-1}{n}\frac{1}{n-1}\sum_{k=1}^{n-1}(z_k z_k' - \Sigma) + \frac{1}{n}\Sigma \xrightarrow{d} N_{p(p+1)/2}(0, V).$$

其中 V 的元素为 $cov(z_{1i}z_{1j}, z_{1s}z_{1t})$. 最后再改写成矩阵形式即证. □

1.3 Wishart 分布

样本协方差矩阵 S 的分布和所谓的 Wishart 分布有关:

Definition

设 $X_k, k = 1\ldots, n$ 为相互独立且服从 $\sim N_p(\mu_k, \Sigma)$ 的随机向量, 则称

$$W = \sum_{i=1}^{n} X_i X_i'$$

的分布为自由度 n 的非中心 **Wishart** 分布. 记作 $W \sim W_p(n, \Sigma, \Delta), \Delta = M'M$ 其中 $M = [\mu_1, \ldots, \mu_n]'$. 当 $M = 0$ 时, 称 W 服从中心 **Wishart** 分布, 通常记作 $W \sim W_p(n, \Sigma)$.

若记 $\mathbf{X} = [X_1, \ldots, X_n]'$, 则 $\mathbf{X} \sim N_{n\times p}(M, I_n \otimes \Sigma)$, $W = \mathbf{X}'\mathbf{X} \sim W_p(n, \Sigma, \Delta)$.

当 $p = 1$ 时, Wishart 分布退化为卡方分布.

在此定义下

定理 6. 设 $X_1, \ldots, X_n i.i.d \sim N_p(\mu, \Sigma)$, S 为样本协方差矩阵, 则

$$(n-1)S \sim W_p(n-1, \Sigma)$$

我们不加证明的列举一些有关 Wishart 分布的性质.

- 当 $n > p, \Sigma > 0$ 时候, $W_p(n, \Sigma)$ 有概率密度函数

$$f(W) = c|W|^{(n-p-1)/2} etr[-\frac{1}{2}\Sigma^{-1}W]I(W > 0).$$

- 若 $W \sim W_p(n, \Sigma)$, 则

$$EW = n\Sigma, Cov(vec(W)) = n(I_{p^2} + K)(\Sigma \otimes \Sigma)$$

其中 $K = \sum_{i,j=1}^{p}(E_{ij} \otimes E'_{ij})$.

- 设 $W \sim W_p(n, \Sigma, \Delta)$, B 为 $q \times p$ 矩阵, 则

$$BWB' \sim W_q(n, B\Sigma B', B\Delta B')$$

- 若 $W_j \sim W_p(n_j, \Sigma, \Delta_j), j = 1, \ldots, m$ 且相互独立, 则

$$\sum_{j=1}^{m} W_j \sim W_p(n, \Sigma, \Delta)$$

其中 $n = \sum_{j=1}^{m} n_j, \Delta = \sum_{j=1}^{m} \Delta_j$.

- 若 $W \sim W_p(n, \Sigma, \Delta)$, 按同样方式分块 W, Σ, Δ:

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

W_{11} 为 $q \times q$ 矩阵, 则

$$W_{11} \sim W_q(n, \Sigma_{11}, \Delta_{11}), W_{22} \sim W_{p-q}(n, \Sigma_{22}, \Delta_{22})$$

W_{11} 和 W_{22} 相互独立当且仅当 $\Sigma_{12} = 0$.

- (Cochran) 设 $X \sim N_{n\times p}(M, I_p \otimes \Sigma)$,$C, D$ 为 n 阶对称阵, 则

  (1) $X'CX \sim W_p(r, \Sigma, \Delta)$ 当且仅当 $C^2 = C, Rank(C) = r, \Delta = M'CM$.

  (2) $X'CX$ 和 $X'DX$ 相互独立, 当且仅当 $CD = 0$.

- 若 $W \sim W_p(n, I_p)$, 将 W 按照前面的分块, 则

$$W_{22\cdot 1} := W_{22} - W_{21}W_{11}^{-1}W_{12} \sim W_{p-q}(n-q, \Sigma_{22\cdot 1})$$

  其中 $\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, 且 $W_{22\cdot 1}$ 与 W_{11} 相互独立.

- (Bartlett 分解) 若 $W \sim W_p(n, I_p), p < n$, 分解 $W = T'T$, 其中 $T = (t_{ij})$ 是对角元素为正的上三角矩阵, 则

  (1) $\{t_{ij}, i < j\}$ 相互独立; (2) $t_{ij} \sim N(0, 1), i < j$

  (3) $t_{ii}^2 \sim \chi_{n-i+1}^2, i = 1, 2, \ldots, p$.

1.4 评估正态性假设

- 一般来说, 许多多元方法依赖于 $\bar{\mathbf{x}}$ 的分布或者距离

$$n(\bar{\mathbf{x}}-\mu)'S^{-1}(\bar{\mathbf{x}}-\mu)$$

- 大样本理论告诉我们如果简单随机样本 $X_1,\ldots,X_n$ 来自均值为 μ, 协方差为 Σ 的总体, 则

$$\sqrt{n}(\bar{\mathbf{x}}-\mu)\to N_p(0,\Sigma),$$
$$n(\bar{\mathbf{x}}-\mu)'S^{-1}(\bar{\mathbf{x}}-\mu)\to\chi^2_p$$

- 因此, 当推断总体均值时, 如果样本量足够大, 则是否假设总体服从正态分布不是特别重要. 但当样本量较小时, 则我们需要检查样本是否来自正态分布.

- 在高维场合下, 评估正态性假设是比较困难的

- 由于当样本来自多元正态总体时候, 其一维边际, 二维边际以及其他一些样本数字特征应该具有一些特点, 因此
  - 一维边际分布是否为正态?
  - 任何两个变量的散点图是否呈现椭圆形状?
  - 常数密度轮廓线包含的比例是否接近其理论值?
  - 条件分布 $E(X_i|X_j)$ 是否为线性的? 条件方差是否与条件变量无关?
- 即便以上问题我们都没有否定, 我们也不能得出样本来自多元正态分布.
- 一般对是否为多元正态分布, 我们仅仅检查必要条件而不是充分条件是否成立

评估一元正态分布

- 对每个分量检查直方图
- 若 $X \sim N(0, \sigma^2)$, 则

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68,$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95,$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.99$$

  当样本量较大时候, 对每个分量, 使用样本估计 μ, σ 后计算上述各区间的比例

- 使用正态 Q-Q 图检查每个分量, 即使用样本分位数和相应的正态分位数作图, 如果样本来自正态总体, 则图中的点应沿一条直线分布.

- 若 $Z \sim N(0,1)$, 则

$$P(Z \le q_{(i)}) \approx p_{(i)} = \frac{i - .5}{n}$$

因此由样本得到 $x_{(1)}, \ldots, x_{(n)}$, 它们分别是 $(1-.5)/n, \ldots, (n-.5)/n$ 分位数. **R**中对 $n \le 10$ 时使用 $\frac{i-3/8}{n+1/4}$ 近似.

- 使用 $(q_{(1)}, x_{(1)}), \ldots, (q_{(n)}, x_{(n)})$ 作图
- 当数据来自正态分布时候,

$$x_{(i)} \approx \sigma q_{(i)} + \mu$$

因此散点图上点应沿一条直线分布.

- Q-Q 图需要样本量较大, 比如 $n \ge 20$. 对小样本量, 即便总体是正态分布, 其波动也较大.

- 其他一些量化检验方法也可以使用:

– Shapiro-Wilks' Test

$$W = \frac{\sum_{i=1}^{n} a_j (x_{(i)} - \bar{x})(q_{(i)} - \bar{q})}{\sqrt{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2 \sum_{i=1}^{n}(q_{(i)} - \bar{q})^2}}$$

若样本来自正态分布, 则 W 值应该靠近 1.

– Anderson-Darling Test

$$\begin{aligned} A_n^2 &= n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x,\hat{\theta})]^2}{F(x,\hat{\theta})[1 - F(x,\hat{\theta})]} dF(x,\hat{\theta}) \\ &= -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)[log(p_i) + log(1 - p_{n+1-i})] \end{aligned}$$

其中 $p_i = \Phi(\frac{x_{(i)} - \bar{x}}{s})$.

若样本来自正态分布, 则 A^2 值应该比较小.

– Kolmogorov-Smirnov Test

$$D_n = \max(D_n^-, D_n^+)$$

其中 $D_n^- = \max_{1\le i\le n} |p_i - \frac{i-1}{n}|$, $D_n^+ = \max_{1\le i\le n} |p_i - \frac{i}{n}|$,$p_i = \Phi(\frac{x_{(i)}-\bar{x}}{s})$.

若样本来自正态分布, 则 D_n 值应该比较小.

– Cramer-von Mises test 设 $x_1 \le x_2 \le \cdots \le x_n$ 为排序后的样本值, 则检验统计量

$$T = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2$$

其中 F 为零假设分布. 当 T 值过大时候拒绝零假设.

参考 R 包 **nortest**.

评估二元或多元正态分布

注意到如果样本 $\mathbf{x}_1, \ldots, \mathbf{x}_n i.i.d \sim N_p(\mu, \Sigma)$, 则

$$d_i^2 = (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) \sim \chi_p^2$$

因此使用 $(\bar{\mathbf{x}}, S)$ 代替 (μ, Σ) 后记为 $\hat{d}_i^2$, 对较大的 $n - p$(至少 ≥ 25), $\hat{d}_i^2$ 也近似服从 χ_p^2. 从而使用 $\hat{d}_1^2, \ldots, \hat{d}_n^2$ 画卡方 Q-Q 图, 则数据点应该近似在一条直线上.

Chi Square qqplot

- 从小到大排序 $\hat{d}_1^2, \ldots, \hat{d}_n^2$ 为 $\hat{d}_{(1)}^2 \leq \cdots \leq \hat{d}_{(n)}^2$
- 计算概率 $\hat{p} = \frac{i-0.5}{n}$, 然后计算 n 个卡方分位数 $q_{c,p}(\hat{p}_i), i = 1, \ldots, n$.
- 使用 $(\hat{d}_{(i)}, q_{c,p}(\hat{p}_i)), i = 1, \ldots, n$ 作图
- 绘制 45^o 直线

多元正态的量化检验方法

1. 多重检验方法

- 注意到对任何 p 元随机变量 X, X 服从 p 元正态分布当且仅当 X 的任意线性组合 $a'X$ 服从一元正态分布
- 从而可以通过 (随机) 选择一大批向量 a, 然后使用某个量化检验来评估每个一元正态是否成立, 最后再综合起来
- 设我们随机选择 N 个独立的单位向量 a, 使用 Shapiro-Wilks' test 评估每个 $a'X$ 的正态性, 最后将 N 个 p 值综合起来
- 注意到我们有 N 个假设需要检验, 因此需要使用多重检验方法. 这里以控制错误发现率 (False Discovery Rate, FDR) 为例:
  - 记所有 N 个假设 $H_0 : a_i'X \sim N(0,1) \leftrightarrow H_1 : a_i'X \not\sim N(0,1), i = 1, \ldots, N$ 的 p 值为 $p_1, \ldots, p_N$.

– FDR 是一种控制一型错误的方法:

$$FDR = E\left[\frac{V}{R}\middle| R > 0\right] P(R > 0)$$

其中 R 表示 N 个检验中拒绝的个数, V 为错误拒绝 (false positive) 个数.

– 欲使 $FDR \leq \alpha$, Benjamini& Hochberg 提出一种方法

1. 将所有 p 值从小到大排序为 $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$;
2. 计算 $\hat{k} = \max\{k : p_{(k)} \leq \alpha k/m\}$;
3. 最后拒绝 $p_{(1)}, \ldots, p_{(\hat{k})}$ 对应的零假设.

当 $\min_{1 \leq i \leq N}\{p_{(i)}m/i\} > \alpha$ 时候, 我们就接受 X 服从正态分布这一零假设.

2. Energy Test(Szekely and Rizzo 2013,JSPI)

- (Energy distance) 两个 p 元随机变量之间的 *energy distance*定义为

$$\mathcal{E}(X,Y) = 2E\|X-Y\| - E\|X-X'\| - E\|Y-Y'\|$$

其中 X, X' $i.i.d$, Y, Y' $i.i.d$.$\|\cdot\|$ 表示欧式模.

- $\mathcal{E}(X,Y) \geq 0$, 等号成立当且仅当 X 和 Y 同分布

- 设 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ 为来自某个 p 元总体分布的标准化样本, 则考虑所谓的 "energy" 检验统计量

$$\hat{\mathcal{E}}_{n,p} = n\left[\frac{2}{n}\sum_{i=1}^{n} E\|\mathbf{x}_i^* - Z\| - E\|Z-Z'\| - \frac{1}{n^2}\sum_{j,k=1}^{n}\|\mathbf{x}_j^* - \mathbf{x}_k^*\|\right]$$

其中 Z, Z' $i.i.d \sim N_p(0, I_p)$. $\mathbf{x}_i^*$ 为使用样本均值和样本协方差标准化后的样本.

- 在 R 中, 可以通过使用包 **energy**中的函数 **mvnorm.etest**进行多元正态检验
- 通过参数 bootstrap 方法确定检验的临界值

1.5 异常点检测

- 异常点 (outlier) 即为与相邻观测点差异巨大的观测点
- 在一元场合, 如果样本量 n 充足且假设总体为正态分布, 则可以通过如下检查异常点 (boxplot)
  - 标准化 n 个样本点
  - 超出 ± 3.5 之外的点可以认为是异常点
- 在 p 维场合, 异常点的检测并不容易. 一个样本点在任意一维边际分布中可能不是异常点, 但在 p 元联合分布下是异常点. 一般的做法
  - 将样本进行标准化, 视觉检查所有的一维边际分布
  - 如果 p 适当, 则可以检查所有的 2 元散点图.

– 对每个样本点 $\mathbf{x}_i$, 计算平方距离

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

由于 d_i^2 近似服从 χ_p^2, 因此当 n 较大时候 (如 $n = 100$), d_i^2 超过 $5q_{c,p}(0.05)$ 时即可认为是异常点.

量化检验是否存在异常点

- Mardia (1970,1974,1980) 定义了多元峰度系数

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} [(\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})]^2$$

- Schwager and Margolin (1982) 表明当 $b_{2,p}$ 大于某个临界值时候, 可以断言样本中存在异常点.
- 因此 p 值 $= P(b_{2,p} \geq b_{2,p,obs} | H_0)$, 其中 H_0 为样本来自标准正态分布.

– 在 H_0 下可以证明

$$\sqrt{n}\frac{b_{2,p}-p(p+2)}{\sqrt{8p(p+2)}} \xrightarrow{d} N(0,1)$$

从而对较大的 n, p 值 $\approx 1-\Phi(\sqrt{n}\frac{b_{2,p,obs}-p(p+2)}{\sqrt{8p(p+2)}})$

– Bootstrap 方法. 从标准 p 元正态分布中产生 n 个样本点 Z, 计算 $b_{2,p}(Z)$, 重复这样 B 次, 得到 H_0 下 $b_{2,p}$ 的经验分布, 于是 p 值 $\approx \#\{b_{2,p}(Z) \geq b_{2,p(\mathbf{x})}\}/B$.

- 这种方法不能得出具体哪个样本点是异常点. 但可以通过检查每个 $(\mathbf{x}_i-\bar{\mathbf{x}})'S^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})$ 来确定是否为异常点.

- **R**的包 **mvoutlier**提供了基于稳健方法的异常点检测手段.

1.6 正态化变换

- 如果观测点明显偏离正态分布, 则可能需要对样本进行必要的变换, 使其近似服从正态分布.

- 对一维数据, 常用的正态化变换有

数据尺度	变换
右偏数据	$log(x)$
计数数据 x	$\sqrt{x}$
x 为百分比数据 p	$logit(p) = \frac{1}{2}log(p/(1-p))$
x 为相关系数 r	Fisher's $z(r) = \frac{1}{2}log[(1+r)/(1-r)]$

- (Box-Cox 变换) Box-Cox (1964, JRSSB) 对连续型数据提出一类变换:

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases}$$

- 注意 x 需要是正值.
- 为估计 λ, 假设变换后的变量 $x^{(\lambda)}$ 服从正态分布 $N(\mu, \sigma^2)$. 则有样本 $x_1, \ldots, x_n$ 后可以通过最大似然方法估计 λ. 由 x 的对数似然函数为

$$l(\mu, \sigma^2, \lambda) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i^{(\lambda)} - \mu)^2 + (\lambda - 1) log(x_i)$$

使用 μ 和 σ^2 的 MLE 代替, 则得到 λ 的 (偏似然) 函数

$$l(\lambda) = -\frac{n}{2} [\frac{1}{n} \sum_{i} (x_{i(\lambda)} - \overline{x^{(\lambda)}})^2] + (\lambda - 1) log(x_i)$$

最好的变换 $\hat{\lambda} = \arg\max_{\lambda > 0} l(\lambda)$.

- 实际使用时在取某个区间 $\lambda \in [a, b]$ 中的一个序列, 在每个 λ 处计算 $l(\lambda)$, 然后找出最小值对应的 λ.

- 对 p 维数据, 在每个一维边际上单独实施 Box-Cox 变换, 这样可以使得一维边际近似正态, 但联合分布不能保证近似正态

- 多元 Box-Cox 变换: 对每个一维实施一元 Box-Cox 变换, 使得变换后的样本服从多元正态分布, 则可以通过最大化对数似然函数来确定 p 个 λ.

- 参考 **R**包 **car**里的函数 **powerTransform**.