(فصل دوازدهم را ملاحظه کنید) باشند . علاوه بر این مؤلفه های اصلی (مقیاس بندی) شده یک «عامل سازی» ماتریس کوواریانس برای الگوی تحلیل عاملی مورد نظر در فصل نهم است .

۸-۲ مؤلفه های اصلی جامعه

مؤلفه های اصلی از نظرجبری ترکیبات خطی ویژهٔ p متغیر تصادفی $X_1, X_2, \ldots, X_p$ است . این ترکیبات خطی از نظر هندسی انتخاب یک دستگاه مختصات جدید را نشان می دهد که از دوران دستگاه اولیه با $X_1, X_2, \ldots, X_p$ به عنوان محورهای مختصات به دست می آید . محورهای جدید جهتها را با بیشترین تغییرپذیری نشان می دهد و بیان ساده تر و ممسک تری از ساختمان کوواریانس را فراهم می کند .

چنان که ملاحظه خواهیم نمود مؤلفه های اصلی تنها به ماتریس کوواریانس $\boldsymbol{\Sigma}$ (یا ماتریس همبستگی $\boldsymbol{\rho}$) $X_1, X_2, \ldots, X_p$ مربوط می شود . برای بسط آنها فرض نرمال چندمتغیری لازم نیست . از سوی دیگر مؤلفه های اصلی که برای جامعه های نرمال چندمتغیری به دست می آید تعابیر مفیدی برحسب بیضویهای چگالی ثابت دارد . علاوه بر این وقتی جامعهٔ نرمال چندمتغیری است (بخش ۸-۵ را ملاحظه نمایید) استنباطهایی را از مؤلفه های نمونه می توان به عمل آورد .

فرض کنید بردار تصادفی $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ دارای ماتریس کوواریانس Σ با مقادیر ویژهٔ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ است .

ترکیبات خطی زیر را در نظر می گیریم :

$$\begin{aligned} Y_1 &= \boldsymbol{\ell}_1'\mathbf{X} = \ell_{11}X_1 + \ell_{21}X_2 + \cdots + \ell_{p1}X_p \\ Y_2 &= \boldsymbol{\ell}_2'\mathbf{X} = \ell_{12}X_1 + \ell_{22}X_2 + \cdots + \ell_{p2}X_p \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ Y_p &= \boldsymbol{\ell}_p'\mathbf{X} = \ell_{1p}X_1 + \ell_{2p}X_2 + \cdots + \ell_{pp}X_p \end{aligned} \tag{۸-۱}$$

در این صورت با استفاده از (۲-۴۵) ، داریم :

$$\text{Var}(Y_i) = \boldsymbol{\ell}_i'\boldsymbol{\Sigma}\boldsymbol{\ell}_i \qquad i = 1, 2, \ldots, p \tag{۸-۲}$$

$$\text{Cov}(Y_i, Y_k) = \boldsymbol{\ell}_i'\boldsymbol{\Sigma}\boldsymbol{\ell}_k \qquad i, k = 1, 2, \ldots, p \tag{۸-۳}$$

مؤلفه های اصلی آن ترکیبات خطی ***ناهمبسته*** $Y_1, Y_2, \ldots, Y_p$ هستند که واریانسهای آنها در (۸-۲) تا جایی که ممکن است ، بزرگ باشد .

اولین مؤلفهٔ اصلی یک ترکیب خطی با واریانس ماکزیمم است . یعنی $\mathrm{Var}(Y_1) = \boldsymbol{\ell}_1'\boldsymbol{\Sigma}\boldsymbol{\ell}_1$ را ماکزیمم می کند . واضح است که $\mathrm{Var}(Y_1) = \boldsymbol{\ell}_1'\boldsymbol{\Sigma}\boldsymbol{\ell}_1$ را می توان با ضرب کردن هر l_1 در یک ثابت ماکزیمم کرد . برای از بین بردن این ابهام بهتر است بردارهای ضرایب با طول واحد را مورد توجه قرار دهیم . بنابراین تعریف می کنیم ،

ترکیب خطی $l'_1\mathbf{X}$ که $\mathrm{Var}(l'_1\mathbf{X})$ را با توجه به $l'_1l_1 = 1$ ماکزیمم کند = اولین مؤلفهٔ اصلی .

ترکیب خطی $l'_2\mathbf{X}$ که $\mathrm{Var}(l'_2\mathbf{X})$ را با توجه به $l'_2l_2 = 1$ و $\mathrm{Cov}(l'_1\mathbf{X}, l'_2\mathbf{X}) = 0$ = دومین مؤلفهٔ اصلی ماکزیمم کند .

در مرحلهٔ i ام

ترکیب خطی $l'_i\mathbf{X}$ که $\mathrm{Var}(l'_i\mathbf{X})$ را با توجه به $\boldsymbol{\ell}_i'\boldsymbol{\ell}_i = 1$ و = مؤلفهٔ اصلی i ام

$\mathrm{Cov}(l'_1\mathbf{X}, l'_k\mathbf{X}) = 0$ ، $k < i$ ماکزیمم کند .

نتیجهٔ ۸-۱ . فرض کنید $\boldsymbol{\Sigma}$ ماتریس کوواریانس بردار تصادفی $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ باشد . فرض کنید Σ دارای زوج مقدار ویژه ـ بردار ویژه $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ باشد که $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ ، ***مؤلفهٔ اصلی*** i ام با

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{1i}X_1 + e_{2i}X_2 + \cdots + e_{pi}X_p, \qquad i = 1, 2, \ldots, p \tag{۸-۴}$$

داده می شود . با این انتخابها ، داریم :

$$\begin{aligned} \mathrm{Var}(Y_i) &= \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i & i &= 1, 2, \ldots, p \\ \mathrm{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_k = 0 & i &\neq k \end{aligned} \tag{۸-۵}$$

در صورتی که بعضی از λ_iها برابر باشند ، انتخابهای بردارهای ضرایب مربوط $\mathbf{e}_i$ و در نتیجه Y_i یکتا نخواهند بود .

اثبات . از (۲-۵۱) با $\mathbf{B} = \boldsymbol{\Sigma}$ می دانیم که

$$\max_{\boldsymbol{\ell} \neq 0} \frac{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}}{\boldsymbol{\ell}'\boldsymbol{\ell}} = \lambda_1 \quad (\text{وقتی به دست می آید که } \boldsymbol{\ell} = \mathbf{e}_1)$$

اما چون بردارهای ویژه نرمال شده اند ، لذا $\mathbf{e}'_1\mathbf{e}_1 = 1$. از این رو

$$\max_{\boldsymbol{\ell} \neq 0} \frac{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}}{\boldsymbol{\ell}'\boldsymbol{\ell}} = \lambda_1 = \frac{\mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{e}_1}{\mathbf{e}_1'\mathbf{e}_1} = \mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{e}_1 = \mathrm{Var}(Y_1)$$

به طور مشابه با استفاده از (۲-۵۲) ، داریم :

$$\max_{\boldsymbol{\ell} \perp \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k} \frac{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}}{\boldsymbol{\ell}'\boldsymbol{\ell}} = \lambda_{k+1} \qquad k = 1, 2, \ldots, p-1$$

برای انتخاب $l = e_{k+1}$ که $\mathbf{e}'_{k+1}\mathbf{e}_k = 0$ ، $k = 1, 2, \ldots, p-1$ ،

$$\mathbf{e}'_{k+1}\boldsymbol{\Sigma}\mathbf{e}_{k+1}/\mathbf{e}'_{k+1}\mathbf{e}_{k+1} = \mathbf{e}'_{k+1}\boldsymbol{\Sigma}\mathbf{e}_{k+1} = \text{Var}(Y_{k+1})$$

اما $\mathbf{e}'_{k+1}(\Sigma\mathbf{e}_{k+1}) = \lambda_{k+1}\mathbf{e}'_{k+1}\mathbf{e}_{k+1} = \lambda_{k+1}$ بنابراین : $\text{Var}(Y_{k+1}) = \lambda_{k+1}$. حال باید عمود بودن $\mathbf{e}_i$ بر $\mathbf{e}_k$ (یعنی $\mathbf{e}'_i\mathbf{e}_k = 0$ ، $i \neq k$) که $\text{Cov}(Y_i, Y_k) = 0$ را می دهد اثبات کنیم . اکنون بردارهای ویژهٔ Σ درصورتی که مقادیر ویژه $\lambda_1, \lambda_2, \ldots, \lambda_p$ متمایز باشند ، متعامدند . اگر تمام مقادیر ویژه متمایز نباشند ، در آن صورت بردارهای ویژهٔ متناظر با مقادیر ویژهٔ مشترك را می توان متعامد انتخاب کرد . بنابراین برای هر دو بردار ویژهٔ $\mathbf{e}_i$ و $\mathbf{e}_k$ ، $\mathbf{e}'_i\mathbf{e}_k = 0$ ، $i \neq k$. چون $\Sigma\mathbf{e}_k = \lambda_k\mathbf{e}_k$ لذا اگر آن را از سمت چپ در $\mathbf{e}'_i$ ضرب کنیم ، برای هر $i \neq k$ ، داریم :

$$\text{Cov}(Y_i, Y_k) = \mathbf{e}'_i\boldsymbol{\Sigma}\mathbf{e}_k = \mathbf{e}'_i\lambda_k\mathbf{e}_k = \lambda_k\mathbf{e}'_i\mathbf{e}_k = 0$$

و بدین ترتیب اثبات کامل می شود .

از نتیجهٔ (۸-۱) معلوم می شود که مؤلفه های اصلی ناهمبسته بوده و واریانس آنها برابر مقادیر ویژهٔ $\boldsymbol{\Sigma}$ است .

نتیجهٔ ۸-۲ . فرض کنید $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ دارای ماتریس کوواریانس Σ با زوج مقدار ویژه-بردار ویژه $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ ، $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ است . فرض کنید $Y_1 = \mathbf{e}'_1\mathbf{X}, Y_2 = \mathbf{e}'_2\mathbf{X}, \ldots, Y_p = \mathbf{e}'_p\mathbf{X}$ مؤلفه های اصلی باشند ، آن گاه

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i)$$

اثبات . از تعریف ۲ الف . ۲۸ داریم ، $\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \text{tr}(\boldsymbol{\Sigma})$ از (۲-۲۰) با $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}$ می توان نوشت $\boldsymbol{\Sigma} = \mathbf{P\Lambda P}'$ که Λ ماتریس قطری مقادیر ویژه و $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$ چنان است که $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$. با استفاده از نتیجهٔ ۲ الف . ۱۲ (ج) ، داریم :

$$\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\mathbf{P\Lambda P}') = \text{tr}(\mathbf{\Lambda P}'\mathbf{P}) = \text{tr}(\boldsymbol{\Lambda}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

از این رو

$$\sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \text{Var}(Y_i)$$

نتیجهٔ ۸-۲ بیان می کند که

$$\text{کل واریانس جامعه} = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p \qquad (۸-۶)$$

و در نتیجه نسبت واریانس کل مربوط به (بیان شده با) مؤلفهٔ اصلی k ام ، عبارت است از :

$$\left(\begin{array}{c}\text{سهم کل واریانس}\\ \text{جامعه مربوط}\\ \text{به مؤلفهٔ اصلی}\\ k\text{ ام}\end{array}\right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p \tag{۷-۸}$$

اگر برای pبزرگ بیشتر واریانس کل جامعه (مثلاً ۸۰ تا ۹۰ درصد) آن را بتوان به سه مؤلفهٔ اول نسبت داد در آن صورت این مؤلفه ها را بدون این که اطلاعات زیادی را از دست دهیم ، می توان «جایگزین» p متغیر اولیه کرد .

هر مؤلفه بردار ضرایب $\mathbf{e}'_i = [e_{1i}, \ldots, e_{ki}, \ldots, e_{pi}]$ نیز ارزش بررسی را دارد . مقدار e_{ki} اهمیت متغیر k ام را در مؤلفهٔ اصلی i ام صرف نظر از متغیرهای دیگر اندازه می گیرد . به ویژه e_{ki} با ضریب همبستگی Y_i و X_k متناسب است .

نتیجهٔ ۸-۳ . اگر $Y_1 = \mathbf{e}'_1\mathbf{X}, Y_2 = \mathbf{e}'_2\mathbf{X}, \ldots, Y_p = \mathbf{e}'_p\mathbf{X}$، مؤلفه های اصلی به دست آمده از ماتریس کوواریانس Σ باشد آن گاه

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p \tag{۸-۸}$$

ضرایب همبستگی بین مؤلفه های Y_i ومتغیرهای X_kاست . دراین جا $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ زوجهای مقدار ویژه ـ بردار ویژه Σ هستند .

اثبات . قرار می دهیم $l'_k = [0, \ldots, 0, 1, 0, \ldots, 0]$ به طوری که $X_k = l'_k\mathbf{X}$ و بنا به (۲-۴۵) ، $\operatorname{Cov}(X_k, Y_i) = \operatorname{Cov}(l'_k\mathbf{X}, \mathbf{e}'_i\mathbf{X}) = l'_k\Sigma\mathbf{e}_i$. چون $\Sigma\mathbf{e}_i = \lambda_i\mathbf{e}_i$ لذا $\operatorname{Cov}(X_k, Y_i) = l'_k\lambda_i\mathbf{e}_i = \lambda_i e_{ki}$ در این صورت $\operatorname{Var}(Y_i) = \lambda_i$ ((۸-۵) را ملاحظه نمایید) و $\operatorname{Var}(X_k) = \sigma_{kk}$ ،

$$\rho_{Y_i, X_k} = \frac{\operatorname{Cov}(Y_i, X_k)}{\sqrt{\operatorname{Var}(Y_i)}\sqrt{\operatorname{Var}(X_k)}} = \frac{\lambda_i e_{ki}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p$$

را حاصل می کند .

مثال فرضی زیر مفاد نتایج ۸-۱ ، ۸-۲ و ۸-۳ را تشریح می کند .

مثال ۸-۱

فرض کنید متغیرهای تصادفی X_1 ، X_2 و X_3 دارای ماتریس کوواریانس زیر باشند :

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

زوجهای مقدار ویژه ـ بردار ویژه ، عبارتند از :

$$\begin{aligned} \lambda_1 &= 5.83, & \mathbf{e}_1' &= [.383,\ -.924,\ 0] \\ \lambda_2 &= 2.00, & \mathbf{e}_2' &= [0,\ 0,\ 1] \\ \lambda_3 &= 0.17, & \mathbf{e}_3' &= [.924,\ .383,\ 0] \end{aligned}$$

بنابراین مؤلفه های اصلی به صورت زیر خواهند بود :

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = .383X_1 - .924X_2 \\ Y_2 &= \mathbf{e}_2'\mathbf{X} = X_3 \\ Y_3 &= \mathbf{e}_3'\mathbf{X} = .924X_1 + .383X_2 \end{aligned}$$

متغیر X_3 یکی از مؤلفه های اصلی است ، زیرا با دو متغیر دیگر ناهمبسته است .

معادلهٔ (۸-۵) را با توجه به اصول اولیه می توان تشریح کرد . برای مثال

$$\begin{aligned} \mathrm{Var}(Y_1) &= \mathrm{Var}(.383X_1 - .924X_2) \\ &= (.383)^2\,\mathrm{Var}(X_1) + (-.924)^2\,\mathrm{Var}(X_2) + 2(.383)(-.924)\,\mathrm{Cov}(X_1, X_2) \\ &= .147(1) + .854(5) - .708(-2) \\ &= 5.83 = \lambda_1 \\ \mathrm{Cov}(Y_1, Y_2) &= \mathrm{Cov}(.383X_1 - .924X_2, X_3) \\ &= .383\,\mathrm{Cov}(X_1, X_3) - .924\,\mathrm{Cov}(X_2, X_3) \\ &= .383(0) - .924(0) = 0 \end{aligned}$$

همچنین به آسانی می توان دید

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + .17$$

که معادلهٔ (۸-۶) را برای این مثال تأیید می کند . سهم کل واریانس که به وسیلهٔ اولین مؤلفهٔ اصلی به حساب می آید $\frac{\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)} = \frac{5.83}{8} = .73$ است . با ارائه این عمل سهم واریانس جامعه که با دو مؤلفه اول به حساب می آید $\frac{(5.83+2)}{8} = .98$ است . در این حالت بدون این که کمترین اطلاعی را از دست دهیم می توانیم مؤلفه های $\cancel{Y_1}$ و Y_2 را جانشین سه متغیر اولیه کنیم .

بالاخره با استفاده از (۸-۸) ، داریم :

$$\rho_{Y_1, X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{.383\sqrt{5.83}}{\sqrt{1}} = .925$$

$$\rho_{Y_1, X_2} = \frac{e_{21}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-.924\sqrt{5.83}}{\sqrt{5}} = -.998$$

نتیجه می گیریم که اهمیت هر یک از متغیرهای X_1 و X_2 در اولین مؤلفهٔ اصلی تقریباً یکسان است . همچنین

$$\rho_{Y_2,X_3} = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1 \quad \text{و} \quad \rho_{Y_2,X_1} = \rho_{Y_2,X_2} = 0 \quad \text{(که باید چنین باشد)}$$

چون مؤلفه سوم بی اهمیت است ، لذا از بقیهٔ همبستگیها می توان صرف نظر کرد .

بررسی مؤلفه های اصلی به دست آمده از متغیرهای تصادفی نرمال چندمتغیری مفید است . فرض کنید $\mathbf{X}$ دارای توزیع $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ باشد . از (۴-۷) می دانیم که بیضویهای متمرکزشده در $\boldsymbol{\mu}$ با چگالی ثابت

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

دارای محورهای $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$ ، $i = 1, 2, ..., p$ بوده که $(\lambda_i, \mathbf{e}_i)$ زوجهای مقدار ویژه-بردار ویژه $\boldsymbol{\Sigma}$ است . نقطه ای که روی محور i ام بیضوی قراردارد دارای مختصاتی متناسب با $\mathbf{e}'_i = [e_{1i}, e_{2i}, ..., e_{pi}]$ در دستگاه مختصات با مبدأ $\boldsymbol{\mu}$ و محورهای $x_1, x_2, ..., x_p$ است . در بحث بعدی[1] بهتر است $\boldsymbol{\mu} = \mathbf{0}$ قرار دهیم .

با توجه به بحث بخش ۲-۳ که $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ ، می توان نوشت :

$$c^2 = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \frac{1}{\lambda_1}(\mathbf{e}'_1\mathbf{x})^2 + \frac{1}{\lambda_2}(\mathbf{e}'_2\mathbf{x})^2 + \cdots + \frac{1}{\lambda_p}(\mathbf{e}'_p\mathbf{x})^2$$

که در آن $\mathbf{e}'_1\mathbf{x}$ ، $\mathbf{e}'_2\mathbf{x}$ ، . . . ، $\mathbf{e}'_p\mathbf{x}$ را به عنوان مؤلفه های اصلی $\mathbf{x}$ می شناسیم . با قرار دادن $y_1 = \mathbf{e}'_1\mathbf{x}$ ، $y_2 = \mathbf{e}'_2\mathbf{x}$ ، . . . ، $y_p = \mathbf{e}'_p\mathbf{x}$ ، داریم :

$$c^2 = \frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \cdots + \frac{1}{\lambda_p}y_p^2$$

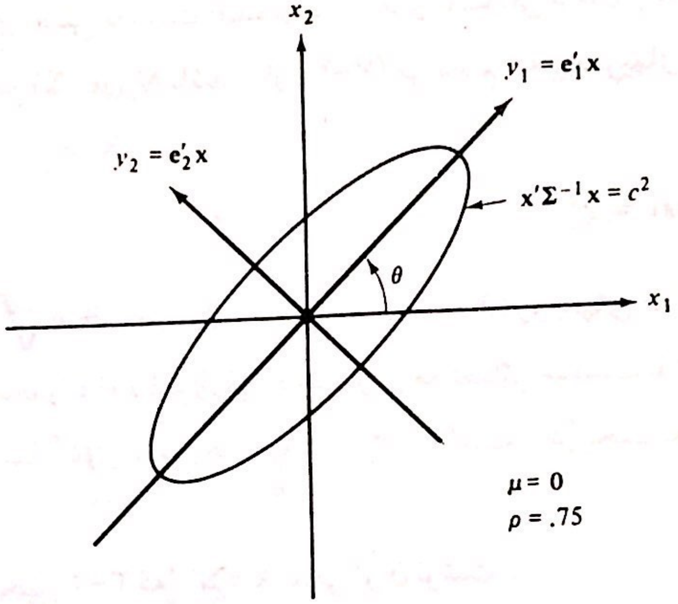
که این معادله یک بیضوی (زیرا $\lambda_1, \lambda_2, ..., \lambda_p$ مثبت اند) را در یک دستگاه مختصات تعریف می کند که محورهای $y_1, y_2, ..., y_p$ آن به ترتیب در جهات $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p$ قرار دارد . اگر λ_1 بزرگترین مقدار ویژه باشد ، آن گاه محور بزرگ آن در جهت $\mathbf{e}_1$ واقع است و محورهای کوچک باقی مانده در جهاتی که با $\mathbf{e}_2, ..., \mathbf{e}_p$ تعریف شده اند ، قرار دارند .

به طور خلاصه مؤلفه های اصلی $y_1 = \mathbf{e}'_1\mathbf{x}$ ، $y_2 = \mathbf{e}'_2\mathbf{x}$ ، . . . ، $y_p = \mathbf{e}'_p\mathbf{x}$ در جهات محورهای بیضوی چگالی ثابت واقع اند . بنابراین هر نقطهٔ روی محور بیضوی i ام دارای مختصات $\mathbf{x}$ متناسب

۱- این را بدون این که به کلیت خللی وارد شود می توان انجام داد ، زیرا بردار تصادفی نرمال $\mathbf{X}$ را همیشه می توان به بردار تصادفی نرمال $\mathbf{W} = \mathbf{X} - \boldsymbol{\mu}$ و $E(\mathbf{W}) = \mathbf{0}$ تبدیل کرد . با وجود این $\mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{W})$.

با $\mathbf{e}'_i = [e_{1i}, e_{2i}, \ldots, e_{pi}]$ و الزاماً مختصات مؤلفهٔ اصلی به شکل $[0, \ldots, 0, y_p, 0, \ldots, 0]$ است.

یک بیضی چگالی ثابت و مؤلفه های اصلی مربوط به بردار تصادفی نرمال دومتغیری با $\boldsymbol{\mu} = \mathbf{0}$ و $\rho = .75$ را در شکل ۸-۱ نشان داده ایم. می بینیم که مؤلفه های اصلی از دوران محورهای مختصات اولیه به اندازه زاویهٔ θ تا این که بر محورهای بیضی چگالی ثابت منطبق شوند، به دست می آیند. این نتیجه برای ابعاد $p > 2$ نیز برقرار است.



شکل ۸-۱ بیضی چگالی ثابت $\mathbf{x}'\Sigma^{-1}\mathbf{x} = c^2$ و مؤلفه‌های اصلی y_1، y_2 برای یك بردار تصادفی نرمال دومتغیره $\mathbf{X}$

محاسبهٔ مؤلفه های اصلی متغیرهای استاندارد شده

مؤلفه های اصلی را می توان برای متغیرهای استاندارد شده نیز به دست آورد:

$$
\begin{aligned}
Z_1 &= \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\
Z_2 &= \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \\
&\vdots \\
Z_p &= \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}
\end{aligned} \qquad (۸-۹)
$$

با نماد ماتریسی می توان نوشت:

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}) \qquad (۸-۱۰)$$

که درآن ماتریس انحراف استاندارد قطری $\mathbf{V}^{1/2}$ را در (۲-۳۵) تعریف کردیم. براساس (۲-۳۷) واضح

است که $E(\mathbf{Z}) = \mathbf{0}$ و

$$\text{Cov}(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

مؤلفه های اصلی $\mathbf{Z}$ را از بردارهای ویژهٔ ماتریس **همبستگی** $\boldsymbol{\rho}$ ، $\mathbf{X}$ می توان به دست آورد . چون واریانس هر Z_i برابر واحد است ، لذا تمام نتایج قبلی را به سهولت می توان به کار برد . ما در آینده نیز از نماد Y_i برای نشان دادن مؤلفهٔ اصلی i ام و $(\lambda_i, \mathbf{e}_i)$ برای زوج مقدار ویژه ـ بردار ویژه استفاده می کنیم . با این وجود ***کمیتهای به دست آمده*** از $\boldsymbol{\Sigma}$ در کل ***همان کمیتهای به دست آمده*** از $\boldsymbol{\rho}$ نیستند .

نتیجهٔ ۸-۴ . مؤلفهٔ اصلی iام متغیرهای استاندارد شدهٔ $\mathbf{Z}' = [Z_1, Z_2, ..., Z_p]$ با $\text{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$ با

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, 2, \ldots, p$$

داده می شود . علاوه بر این

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p \qquad (۱۱-۸)$$

و

$$\rho_{Y_i, Z_k} = e_{ki}\sqrt{\lambda_i}, \qquad i, k = 1, 2, \ldots, p$$

در این حالت $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), ..., (\lambda_p, \mathbf{e}_p)$ زوجهای مقدارویژه ـ بردارویژهٔ $\boldsymbol{\rho}$ با $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ هستند .

اثبات . اگر $Z_1, Z_2, ..., Z_p$ را جایگزین $X_1, X_2, ..., X_p$ کنیم ، $\boldsymbol{\rho}$ را به جای $\boldsymbol{\Sigma}$ قرار دهیم ، نتیجهٔ ۸-۴ از نتایج ۸-۱ ، ۸-۲ و ۸-۳ به دست می آید .

از (۸-۱۱) می بینیم که واریانس کل جامعه (متغیرهای استاندارد شده) برابر p یعنی مجموع اعضای قطری ماتریس $\boldsymbol{\rho}$ است . اگر در (۸-۷) از $\mathbf{Z}$ به جای $\mathbf{X}$ استفاده کنیم سهم واریانس کل بیان شده با مؤلفهٔ اصلی k ام $\mathbf{Z}$ ،

$$\left(\begin{array}{c}\text{سهم واریانس (استاندارد شدهٔ)}\\ \text{جامعه مربوط به مؤلفه اصلی}\\ k\text{ ام}\end{array}\right) = \frac{\lambda_k}{p}, \qquad k = 1, 2, \ldots, p \qquad (۱۲-۸)$$

است که در آن λ_kها مقادیر ویژهٔ $\boldsymbol{\rho}$ هستند .

مثال ۸-۲ (مؤلفه های اصلی حاصل از ماتریسهای کوواریانس همبستگی)
ماتریس کوواریانس

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

و ماتریس همبستگی به دست آمده از آن

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$

را در نظر می گیریم . زوجهای مقدار ویژه-بردار ویژهٔ $\boldsymbol{\Sigma}$ ، عبارتند از :

$$\lambda_1 = 100.16, \qquad \mathbf{e}_1' = [.040, .999]$$
$$\lambda_2 = .84, \qquad \mathbf{e}_2' = [.999, -.040]$$

به طور مشابه زوجهای مقدار ویژه-بردار ویژهٔ حاصل از $\boldsymbol{\rho}$، عبارتند از :

$$\lambda_1 = 1 + \rho = 1.4, \qquad \mathbf{e}_1' = [.707, .707]$$
$$\lambda_2 = 1 - \rho = .6, \qquad \mathbf{e}_2' = [.707, -.707]$$

مؤلفه های اصلی مربوط به صورت زیر در می آید :

$$\boldsymbol{\Sigma}: \quad \begin{aligned} Y_1 &= .040\, X_1 + .999\, X_2 \\ Y_2 &= .999\, X_1 - .040\, X_2 \end{aligned}$$

و

$$\boldsymbol{\rho}: \quad \begin{aligned} Y_1 &= .707Z_1 + .707Z_2 = .707\left(\frac{X_1 - \mu_1}{1}\right) + .707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= .707(X_1 - \mu_1) + .0707(X_2 - \mu_2) \\ Y_2 &= .707Z_1 - .707Z_2 = .707\left(\frac{X_1 - \mu_1}{1}\right) - .707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= .707(X_1 - \mu_1) - .0707(X_2 - \mu_2) \end{aligned}$$

X_2 به خاطر واریانس بزرگی که دارد اولین مؤلفهٔ اصلی تعیین شده از $\boldsymbol{\Sigma}$ را کاملاً تحت تأثیر قرار می دهد . علاوه بر این اولین مؤلفهٔ اصلی یک نسبت

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = .992$$

از واریانس کل جامعه را بیان می کند . در عین حال وقتی متغیرهای X_1 و X_2 را استاندارد می کنیم ، متغیرهای حاصل سهمی یکسان در مؤلفه های اصلی تعیین شده از $\boldsymbol{\rho}$ دارند . با استفاده از ۸-۴ ، داریم :

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$$

و

$$\rho_{Y_1,Z_2} = e_{21}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$$

در این حالت اولین مؤلفهٔ اصلی یک نسبت

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = .7$$

از واریانس کل جامعهٔ (استاندارد شده) را بیان می کند .

ملاحظه می کنیم که اهمیت نسبی متغیرها ، مثلاً روی اولین مؤلفهٔ اصلی به طور قابل ملاحظه ای تحت تأثیر استاندارد کردن واقع می شود . هنگامی که مؤلفه های اصلی به دست آمده از $\boldsymbol{\rho}$ را بر حسب X_1 و X_2 بیان می کنیم ، مقدار نسبی وزنهای .707 و .707 با مقدار نسبی وزنهای .040 و .999 منضم به این متغیرها در مؤلفه های اصلی به دست آمده از $\boldsymbol{\Sigma}$ در جهت مخالف یکدیگرند .

مثال قبلی بیان می کند که مؤلفه های اصلی به دست آمده از $\boldsymbol{\Sigma}$ با مؤلفه های اصلی به دست آمده از $\boldsymbol{\rho}$ متفاوت اند . علاوه بر این یک مجموعه از مؤلفه های اصلی تابعی ساده از سایر مؤلفه ها نیست . که این بدان معنی است که استاندارد کردن بی ربط نیست .

متغیرها را در صورتی که با مقیاسهایی که به طور وسیعی با هم تفاوت دارند، واحدهای اندازه گیری آنها متناسب نیست اندازه گیری کنیم ، احتمالاً باید استاندارد شوند . برای مثال اگر X_1 فروش سالانه را در فاصلهٔ ۱۰٫۰۰۰ دلار تا ۳۵۰٫۰۰۰ دلار نشان دهد و X_2 نسبت درآمد خالص سالانه / کل دارایی که در فاصله ۰٫۰۱ تا ۰٫۶۰ قرار دارد را نشان دهد در آن صورت کل تغییرات تقریباً به خاطر فروش دلار خواهد بود . در این حالت انتظار یک مؤلفهٔ اصلی (مهم) با وزن دار کردن زیاد X_1 را داریم . از طرفی اگر هر دو متغیر را استاندارد کنیم ، مقادیر بعدی آنها به همان ترتیب خواهد بود و X_2 (یا Z_2) نقش مهمتری را در ساختار مؤلفه ها ایفا می کند . که این رفتار را در مثال (۸-۲) مشاهده کردیم .

مؤلفه های اصلی برای ماتریسهای کوواریانس با ساختارهای ویژه

ماتریسهای کوواریانس و همبستگی دارای طرحهای خاصی هستند که مؤلفه های اصلی آنها را به شکلهای ساده می توان بیان کرد . فرض کنید $\boldsymbol{\Sigma}$ ماتریس قطری زیر باشد :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \qquad (۱۳-۸)$$

با قرار دادن $\mathbf{e}'_i = [0, ..., 0, 1, 0, ..., 0]$ که 1 جملهٔ iام آن است ، مشاهده می کنیم که

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1\sigma_{ii} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{یا} \quad \boldsymbol{\Sigma}\mathbf{e}_i = \sigma_{ii}\mathbf{e}_i$$

و نتیجه می گیریم که $(\sigma_{ii}, \mathbf{e}_i)$ زوج مقدار ویژه ـ بردار ویژهٔ i ام است . چون ترکیب خطی $\mathbf{e}'_i\mathbf{X} = X_i$ ، لذا مجموعهٔ مؤلفه های اصلی درست مجموعهٔ اولیه متغیرهای تصادفی ناهمبسته است .

برای یک ماتریس کوواریانس با طرح (۸-۱۳) با خلاصه کردن مؤلفه های اصلی چیزی به دست نمی آید . از نقطهٔ نظر دیگر اگر $\mathbf{X}$ دارای توزیع $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ باشدمسیرهای چگالی ثابت بیضویهایی هستند که محورهای آن قبلاً در جهات بیشترین تغییر بود . بنابراین نیازی به دوران دستگاه مختصات نیست .

استاندارد کردن وضعیت را برای Σ در (۸-۱۳) به طور اساسی تغییر نمی دهد . در این حالت $\boldsymbol{\rho} = \mathbf{I}$ ، ماتریس همانی $p \times p$ است . واضح است که $\boldsymbol{\rho}\mathbf{e}_i = 1\mathbf{e}_i$، لذا مقدار ویژهٔ 1 دارای مضرب p است و $\mathbf{e}'_i = [0, ..., 0, 1, 0, ..., 0]$ ، $i = 1, 2, ..., p$ انتخابهای مناسب بردارهای ویژه هستند . در نتیجه مؤلفه های اصلی که برای $\boldsymbol{\rho}$ به دست می آیند ، مؤلفه های اصلی متغیرهای اولیهٔ $Z_1, Z_2, ..., Z_p$ نیز هستند . علاوه بر این در این حالت که مقادیر ویژه برابرند ، بیضویهای نرمال چندمتغیری چگالی ثابت کروی هستند .

طرح دیگری از ماتریس کوواریانس که اغلب رابطهٔ میان بعضی متغیرهای مربوط به زیست شناسی نظیر میزان حیات اشیاء را بیان می کند ، شکل کلی زیر را دارد :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \tag{۱۴-۸}$$

ماتریس همبستگی که از این ماتریس نتیجه می شودنیزماتریس کوواریانس متغیرهای استانداردشده است .

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{۱۵-۸}$$

از ماتریس (۸-۱۵) نتیجه می شود که متغیر $X_1, X_2, ..., X_p$ همبستگی یکسانی دارند .

اثبات این مطلب که (تمرین ۸-۵ را ملاحظه کنید) p مقدار ویژهٔ ماتریس همبستگی (۸-۱۵) را به دو دسته می توان تقسیم کرد مشکل نیست . وقتی ρ مثبت است بزرگترین مقدار ویژه

$$\lambda_1 = 1 + (p - 1)\rho \qquad (۸\text{-}۱۶)$$

متناظر با بردار ویژهٔ

$$\mathbf{e}_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right] \qquad (۸\text{-}۱۷)$$

است . $p - 1$ مقدار ویژهٔ باقی مانده

$$\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$$

بوده و یک انتخاب بردارهای ویژه ، عبارت است از :

$$\mathbf{e}_2' = \left[\frac{1}{\sqrt{1 \times 2}}, \frac{-1}{\sqrt{1 \times 2}}, 0, \ldots, 0\right]$$

$$\mathbf{e}_3' = \left[\frac{1}{\sqrt{2 \times 3}}, \frac{1}{\sqrt{2 \times 3}}, \frac{-2}{\sqrt{2 \times 3}}, 0, \ldots, 0\right]$$

$$\vdots$$

$$\mathbf{e}_i' = \left[\frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, \ldots, 0\right]$$

$$\vdots$$

$$\mathbf{e}_p' = \left[\frac{1}{\sqrt{(p-1)p}}, \ldots, \frac{1}{\sqrt{(p-1)p}}, \frac{-(p-1)}{\sqrt{(p-1)p}}\right]$$

اولین مؤلفه اصلی

$$Y_1 = \mathbf{e}_1'\mathbf{X} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} X_i$$

با مجموع p متغیر اولیه متناسب است ، که این را می توان به عنوان یک «شاخص» با وزنهای مساوی تلقی کرد . این مؤلفه اصلی یک نسبت

$$\frac{\lambda_1}{p} = \frac{1 + (p-1)\rho}{p} = \rho + \frac{1 - \rho}{p} \qquad (۸\text{-}۱۸)$$

از کل تغییرات جامعه را بیان می کند . ملاحظه می کنیم که برای ρ نزدیک به 1 یا p بزرگ $\rho = \frac{\lambda_1}{p}$. برای مثال اگر $\rho = .80$ و $p = 5$ باشد ، مؤلفهٔ اول 84% کل واریانس را بیان می کند . وقتی ρ نزدیک به 1 است $p - 1$ مؤلفهٔ آخر جمعاً سهم بسیار کمی از کل واریانس را دارند و اغلب از آنها می توان صرف نظر کرد .

اگر متغیرهای استاندارد شدهٔ $Z_1, Z_2, ..., Z_p$ دارای یک توزیع نرمال چندمتغیری با ماتریس کوواریانسی که با (۸-۱۵) داده می شود باشد آن گاه بیضویهای چگالی ثابت «سیگاری شکل» بوده و محور بزرگ (اصلی) در امتداد اولین مؤلفهٔ اصلی $Y_1 = (\frac{1}{p})[1, 1, ..., 1]$ است . محورهای کوچک (فرعی) (و بقیهٔ مؤلفه های اصلی) در جهات شکل کروی متقارن عمود بر محور بزرگ (و مؤلفهٔ اصلی اول) روی می دهند .

۸-۳ خلاصه کردن تغییرات نمونه به وسیلهٔ مؤلفه های اصلی

ما اکنون چارچوبی را که برای مطالعهٔ مسألهٔ خلاصه کردن تغییرات در n راندازه وی p متغیر لازم است را با چند انتخاب درست ترکیبات خطی در دست داریم :

فرض می کنیم داده های $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ استخراج مستقل از جامعهٔ p بُعدی با بردار میانگین $\boldsymbol{\mu}$ و ماتریس واریانس $\boldsymbol{\Sigma}$ را نشان می دهد این داده ها بردار میانگین نمونه $\bar{\mathbf{x}}$ ، ماتریس واریانس نمونهٔ $\mathbf{S}$ و ماتریس همبستگی نمونه $\mathbf{R}$ را حاصل می کند .

هدف ما در این بخش این است که ترکیبات خطی ناهمبسته ای از خصیصه های اندازه گیری شده ای را بسازیم که در بیشتر تغییرات نمونه به حساب می آیند ترکیبات ناهمبسته با بزرگترین واریانس را مؤلفه های اصلی نمونه می نامیم .

خاطر نشان می کنیم که n مقدار هر ترکیب خطی

$$\ell_1'\mathbf{x}_j = \ell_{11}x_{1j} + \ell_{21}x_{2j} + \cdots + \ell_{p1}x_{pj}, \qquad j = 1, 2, \ldots, n$$

دارای میانگین نمونه $\mathbf{l}_1'\bar{\mathbf{x}}$ واریانس نمونهٔ $\mathbf{l}_1'\mathbf{S}\mathbf{l}_1$ است . همچنین زوج مقادیر $(\mathbf{l}_1'\mathbf{x}_j, \mathbf{l}_2'\mathbf{x}_j)$ برای دو ترکیب خطی دارای ماتریس کوواریانس نمونهٔ $\mathbf{l}_2'\mathbf{S}\mathbf{l}_2$ است [(۳-۳۶) را ملاحظه نمایید] .

مؤلفه های اصلی نمونه را به صورت ترکیبات خطی که دارای واریانس نمونه ماکزیمم است ، تعریف می کنیم . بردارهای ضرایب $\mathbf{l}_i$ مانند کمیتهای جامعه باید در $\mathbf{l}_i'\mathbf{l}_i = 1$ صدق کنند ، به ویژه

اولین مؤلفهٔ اصلی **نمونه** = ترکیب خطی $\mathbf{l}_1'\mathbf{x}_j$ که واریانس نمونهٔ $\mathbf{l}_1'\mathbf{x}_j$ را با شرط $\mathbf{l}_1'\mathbf{l}_1 = 1$ ماکزیمم می کند .

دومین مؤلفهٔ اصلی نمونه = ترکیب خطی $l'_2\mathbf{x}_j$ که واریانس نمونهٔ $l'_2\mathbf{x}_j$ را با شرط $l'_1 l_1 = 1$ و کوواریانس صفر برای زوجهای $(l'_1\mathbf{x}_j, l'_2\mathbf{x}_j)$ ماکزیمم کند.

در مرحلهٔ i ام

i امین مؤلفه اصلی = ترکیب خطی $l'_i\mathbf{x}_j$ که واریانس نمونهٔ $l'_i\mathbf{x}_j$ را با شرط $l'_i l_i = 1$ و کوواریانس نمونه صفر برای تمام زوجهای $(l'_i\mathbf{x}_j, l'_k\mathbf{x}_j)$ ، $k < i$ ماکزیمم کند.

اولین مؤلفه اصلی $l'_1 S l_1$ یا معادل با آن

$$\frac{\ell'_1 \mathbf{S} \ell_1}{\ell'_1 \ell_1} \tag{۸-۱۹}$$

را ماکزیمم می کند. بر اساس (۲-۵۱) ماکزیمم آن بزرگترین مقدار ویژهٔ $\hat{\lambda}_i$ است که از انتخاب بردار ویژه $l_1 = \hat{\mathbf{e}}_1$ ، ماتریس S به دست می آید. به شرطی که $\mathbf{l}'_i \mathbf{S} \hat{\mathbf{e}}_k = \mathbf{l}'_i \hat{\lambda}_k \hat{\mathbf{e}}_k = 0$ یا l_i بر $\hat{\mathbf{e}}_k$ عمود باشد انتخابهای متوالی l_i ، (۸-۱۹) را ماکزیمم می کند. بنابراین مانند اثبات نتایج ۸.۱-۸.۳ نتایج زیر که مربوط به مؤلفه های اصلی نمونه است را به دست می آوریم.

اگر $\mathbf{S} = \{S_{ik}\}$ ماتریس واریانس نمونهٔ $p \times p$ با زوجهای مقدار ویژه بردار ویژه $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p \hat{\mathbf{e}}_p)$ باشد، در آن صورت مؤلفهٔ اصلی نمونهٔ i ام با

$$\hat{y}_i = \hat{\mathbf{e}}'_i \mathbf{x} = \hat{e}_{1i}x_1 + \hat{e}_{2i}x_2 + \cdots + \hat{e}_{pi}x_p, \qquad i = 1, 2, \ldots, p$$

داده می شود که $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ و $\mathbf{x}$ هر مشاهدهٔ روی متغیرهای $X_1, X_2, \ldots, X_p$ است.

همچنین

$$\text{واریانس نمونه}(\hat{y}_k) = \hat{\lambda}_k, \qquad k = 1, 2, \ldots, p$$
$$\text{کوواریانس نمونه}(\hat{y}_i, \hat{y}_k) = 0, \qquad i \neq k \tag{۸-۲۰}$$

علاوه بر این

$$\text{کل واریانس نمونه} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

و

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \qquad i, k = 1, 2, \ldots, p$$

ما صرف نظر از این که مؤلفه های اصلی نمونه را از S یا R به دست آورده ایم ، آنها را با $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$ نشان می دهیم . مؤلفه هایی که از S یا R حاصل می شوند در کل یکسان نیستند ولی از فرضیه واضح است که از کدام ماتریس استفاده شده و کدام نماد $\hat{y}_i$ مناسب است . همچنین بهتر است که مؤلفهٔ بردارهای ضرایب را با $\hat{\mathbf{e}}_i$ و مؤلفهٔ واریانسها را برای هر دووضعیت با $\hat{\lambda}_i$ نشان دهیم .

اغلب مشاهدات $\mathbf{x}_j$ را با کم کردن از $\bar{\mathbf{x}}$ «مرکزی» می کنیم . این موضوع روی ماتریس کوواریانس نمونهٔ S تأثیری ندارد و برای هر بردار مشاهدهٔ x ، مؤلفهٔ اصلی زیر را می دهد :

$$\hat{y}_i = \hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}}), \qquad i = 1, 2, \ldots, p \tag{۸-۲۱}$$

اگر مقادیر مؤلفهٔ i ام

$$\hat{y}_{ij} = \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad i = 1, 2, \ldots, p \tag{۸-۲۲}$$

که از جایگزین کردن هر مشاهدهٔ $\mathbf{x}_j$ به جای x دلخواه در (۸-۲۱) تولید می شود ، آن گاه :

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{1}{n}\hat{\mathbf{e}}_i'\left(\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})\right) = \frac{1}{n}\hat{\mathbf{e}}_i'\mathbf{0} = 0 \tag{۸-۲۳}$$

یعنی میانگین نمونهٔ هر مؤلفهٔ اصلی صفر است . واریانسهای نمونه ، مانند (۸-۲۰) هنوز با $\hat{\lambda}_i$ ها داده می شود .

مثال ۸-۳

آمارگیری سال ۱۹۷۰ اطلاعات ناحیه ای روی ۵ متغیر اقتصادی اجتماعی را برای ناحیهٔ

۱- اگر $\mathbf{X}_j$ ها دارای توزیع نرمال باشند (نتیجهٔ ۴-۱۱ را ملاحظه کنید) در آن صورت مؤلفه های اصلی نمونه را می توان از $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_n$ یعنی برآورد درست نمایی ماکزیمم ماتریس کوواریانس $\boldsymbol{\Sigma}$ نیز به دست آورد. در این حالت به شرط این که مقادیر ویژهٔ $\boldsymbol{\Sigma}$ متمایز باشند ، مؤلفه های اصلی نمونه را می توان به عنوان برآوردهای درست نمایی ماکزیمم پارامترهای متناظر جامعهٔ مربوط در نظر نخواهیم گرفت ([۱] را ملاحظه کنید) . چون فرض نرمال در این بخش لازم نیست، لذا $\hat{\boldsymbol{\Sigma}}$ را در نظر گرفت . همچنین $\hat{\boldsymbol{\Sigma}}$ دارای مقادیر ویژهٔ $[(n-1)/n]\hat{\lambda}_i$ و بردارهای ویژهٔ متناظر $\hat{\mathbf{e}}_i$ اند که $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ زوجهای مقدار ویژه-بردار ویژهٔ S هستند . از این رو S و $\hat{\boldsymbol{\Sigma}}$ هر دو مؤلفه های اصلی یکسان $\hat{\mathbf{e}}_i'\mathbf{x}$ [(۸-۲۰) را ملاحظه نمایید] و نسبت واریانس بیان شدهٔ یکسان $\hat{\lambda}_i / (\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_p)$ را می دهد . سرانجام S و $\hat{\boldsymbol{\Sigma}}$ هر دو ماتریس همبستگی نمونهٔ یکسان R را می دهند ، لذا اگر متغیرها را استاندارد کرده باشیم . انتخاب S یا $\hat{\boldsymbol{\Sigma}}$ فرق نمی کند .

مدیسون ویسکانسین فراهم می کند . داده های مربوط به ۱۴ ناحیه را در جدول (۸-۲) در تمرینهای آخرین فصل ثبت کرده ایم . از این داده ها آماره های زیر به دست می آید :

$$\bar{\mathbf{x}}' = [4.32,\quad 14.01,\quad 1.95,\quad 2.17,\quad 2.45]$$

4.32	14.01	1.95	2.17	2.45
کل جامعه (به هزار)	میانهٔ سالهای مدرسه	کل استخدامی (به هزار)	خدمات بهداشتی استخدامی (به صد)	میانهٔ ارزش منزل (\$10,000s)

و

$$\mathbf{S} = \begin{bmatrix} 4.308 & 1.683 & 1.803 & 2.155 & -.253 \\ 1.683 & 1.768 & .588 & .177 & .176 \\ 1.803 & .588 & .801 & 1.065 & -.158 \\ 2.155 & .177 & 1.065 & 1.970 & -.357 \\ -.253 & .176 & -.158 & -.357 & .504 \end{bmatrix}$$

آیا می توان تغییرات را به یک یا دو مؤلفهٔ اصلی خلاصه کرد ؟

جدول زیر را پیدا می کنیم .

ضرایب مربوط به مؤلفه‌های اصلی (اعداد داخل پرانتزها ضرایب همبستگی‌اند)

متغیر	$\hat{e}_1 (r_{\hat{y}_1, x_k})$	$\hat{e}_2 (r_{\hat{y}_2, x_k})$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
کل جامعه	۰٫۷۸۱ (۰٫۹۹)	-۰٫۰۷۱ (-۰٫۰۴)	۰٫۰۰۴	۰٫۵۴۲	-۰٫۳۰۲
میانهٔ سالهای مدرسه	۰٫۳۰۶ (۰٫۶۱)	-۰٫۷۶۴ (-۰٫۷۶)	-۰٫۱۶۲	-۰٫۵۴۵	-۰٫۰۱۰
کل استخدامی	۰٫۳۳۴ (۰٫۹۸)	۰٫۰۸۳ (۰٫۱۲)	۰٫۰۱۵	۰٫۰۵	۰٫۹۳۷
خدمات بهداشتی استخدامی	۰٫۴۲۶ (۰٫۸۰)	۰٫۵۷۹ (۰٫۵۵)	۰٫۲۲۰	-۰٫۶۳۶	-۰٫۱۷۳
میانهٔ ارزش منزل	-۰٫۰۵۴ (-۰٫۲۰)	-۰٫۲۶۲ (-۰٫۴۹)	۰٫۹۶۲	-۰٫۰۵۱	۰٫۰۲۴
پراش ($\hat{\lambda}_i$)	۶٫۶۳۹	۱٫۷۸۶	۰٫۳۹۰	۰٫۲۳۰	۰٫۰۱۴
درصد تجمعی واریانس کل	۷۴٫۱	۹۳٫۲	۹۷٫۴	۹۹٫۹	۱۰۰

اولین مؤلفهٔ اصلی ۷۴٫۱٪ واریانس کل نمونه را بیان می کند . دو مؤلفهٔ اصلی اول با هم ۹۳٫۲٪ واریانس کل را بیان می کنند . در نتیجه تغییرات نمونه خیلی خوب با دو مؤلفهٔ اصلی خلاصه می شود و یک کاهش در داده ها از چهارده مشاهدهٔ روی پنج متغیر به چهارده مشاهدهٔ روی دو مؤلفهٔ اصلی معقول است .

با معلوم بودن ضرایب بالا ، به نظر می رسد مؤلفهٔ اصلی اول یک متوسط موزون چهار متغیر اول است . دومین مؤلفهٔ اصلی به نظر می رسد خدمات سلامتی استخدام را با یک متوسط وزن دار میانهٔ سالهای مدرسه و میانهٔ ارزش منزل مقابله می کند .

هنگامی که می خواهیم در بارهٔ موضوع تعبیر مؤلفه های اصلی کار کنیم ، شاید همبستگیهای $r_{\hat{y}_i, x_k}$ راهنماهای قابل اعتمادتری از مؤلفهٔ ضرایب $\hat{e}_{ki}$ باشند . همبستگیها تفاوتهای در واریانسهای متغیرهای اولیه را مجاز می نمایند که بدین وسیله مشکل تعبیر و تفسیری که مقایسهای متفاوت اندازه گیری موجب آن می شود برطرف می گردد . در مثال (۸-۳) ضرایب همبستگی که در جدول نشان داده شده است ، تعبیری که با مؤلفهٔ ضرایب حاصل می شود را تأیید می کند .

مثال ۸-۴

در مطالعهٔ روابط موجود در اندازه و شکل لاك پشتهای رنگ شده جالیکلو ر و موسیمان [۱۰] درازا ، عرض و ارتفاع لاك پشت را اندازه گیری نموده اند . این داده ها در تمرین (۶ . ۱۳) آورده شده و جدول (۶-۵) تحلیلی را برحسب لگاریتمها پیشنهاد می کند . (جالیکور معمولاً یک تبدیل لگاریتمی را در مطالعهٔ روابط اندازه و شکل پیشنهاد می کند) . یک تحلیل مؤلفهٔ اصلی را انجام می دهیم .

لگاریتمهای طبیعی ابعاد ۲۴ لاك پشت نیز دارای بردار میانگین نمونه $\bar{\mathbf{x}}' = [4.725, 4.478, 3.703]$ و ماتریس کوواریانس

$$\mathbf{S} = 10^{-3}\begin{bmatrix} 11.555 & 8.367 & 8.508 \\ 8.367 & 6.697 & 6.264 \\ 8.508 & 6.264 & 7.061 \end{bmatrix}$$

است . یک تحلیل مؤلفهٔ اصلی خلاصه زیر را به ما می دهد .

ضرایب مربوط به مؤلفه های اصلی (ضرایب همبستگی در داخل پرانتزها هستند)

متغیر	$\hat{e}_1 (r_{\hat{y}_1, x_k})$	$\hat{e}_2$	$\hat{e}_3$
لگاریتم (درازا)	۰/۶۸۳ (۰/۹۹)	۰/۱۶۲	۰/۷۱۲
لگاریتم (عرض)	۰/۵۱۰ (۰/۹۷)	۰/۵۹۱	-۰/۶۲۴
لگاریتم (ارتفاع)	۰/۵۲۲ (۰/۹۷)	-۰/۷۹۰	-۰/۳۲۱
واریانس ($\hat{\lambda}_i$)	۲۴/۳۱ × ۱۰$^{-3}$	۰/۶۳ × ۱۰$^{-3}$	۰/۳۸ × ۱۰$^{-3}$
درصد تجمعی واریانس کل	۹۶/۰	۹۸/۵	۱۰۰

اولین مؤلفهٔ اصلی که ۹۶٪ واریانس کل را بیان می کند یک تعبیر جالبی دارد . چون

$$\hat{y}_1 = .683 \ln(\text{length}) + .51 \ln(\text{width}) + .522 \ln(\text{height}) = \ln[(\text{طول})^{.683}(\text{عرض})^{.510}(\text{ارتفاع})^{.522}]$$

لذا اولین مؤلفهٔ اصلی را می توان به عنوان (حجم) ln جعبه ای با ابعاد تعدیل شده درنظر گرفت .
داده ها را از نظر هندسی می توان به صورت n نقطه در فضای pـ بُعدی رسم کرد . اگر S معین و مثبت باشد ، آن گاه تمام بردارهای $p \times 1$ ، $\mathbf{x}$ که در

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2 \tag{۸-۲۴}$$

صدق می کند یک ابر بیضوی متمرکز در $\bar{\mathbf{x}}$ که محورهایش با بردارهای ویژهٔ $\mathbf{S}^{-1}$ یا معادل با آن با بردارهای ویژهٔ S داده می شود را تعریف می کند (بخش (۲-۳) و نتیجهٔ (۴-۱) را که S جایگزین $\boldsymbol{\Sigma}$ می شود ، ملاحظه کنید) طولهای این محورها متناسب با $\sqrt{\hat{\lambda}_i}$ ، $i = 1, 2, \ldots, p$ بوده که $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ مقادیر ویژهٔ S هستند .

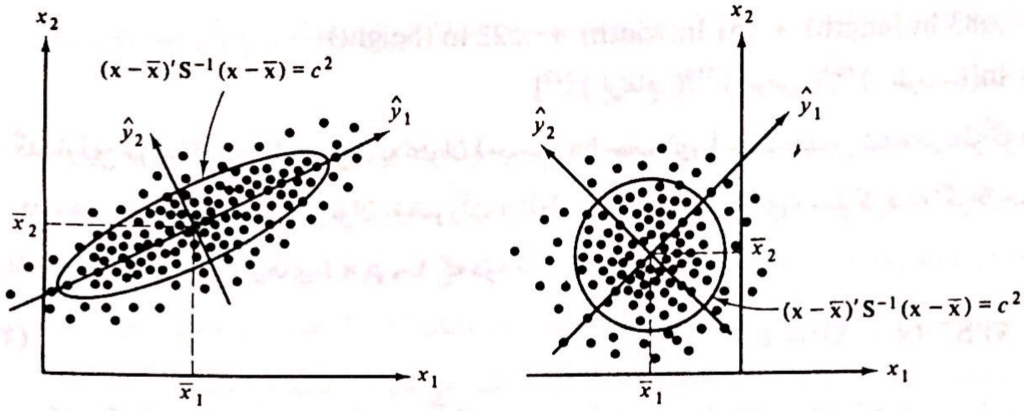
مؤلفه های اصلی نمونه همان ارتباطی را با محورهای بیضویهای با فاصلهٔ ثابت در (۸-۲۴) دارند که مؤلفه های جامعه با بیضویهای با چگالی ثابت برای متغیرهای $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ دارند . یعنی مؤلفه های نمونه در امتداد محورهای بیضویهای با فاصلهٔ ثابت قرار دارد . به این مؤلفه ها می توان به عنوان نتیجهٔ دوران دستگاه مختصات اولیه تا این که محورهای مختصات در جهات واریانس ماکزیمم از نمودار پراکندگی عبور کند ، نگاه کرد .

قدر مطلق مؤلفهٔ اصلی i ام، $|\hat{y}_i| = |\hat{\mathbf{e}}'_i(\mathbf{x} - \bar{\mathbf{x}})|$ طول تصویر بردار $(\mathbf{x} - \bar{\mathbf{x}})$ روی بردار واحد $\hat{\mathbf{e}}_i$ را می دهد [(۲-۸) و (۲-۹) را ملاحظه کنید] . این تعبیر هندسی مؤلفه های اصلی نمونه را در شکل ۸-۲ برای $p = 2$ تشریح نموده ایم .

شکل ۸-۲ (الف) یک بیضی با فاصلهٔ ثابت متمرکز در $\bar{\mathbf{x}}$ را با $\hat{\lambda}_1 > \hat{\lambda}_2$ نشان می دهد . مؤلفه های اصلی نمونه به خوبی تعیین می شوند . این مؤلفه ها در امتداد محورهای بیضی در جهات عمود واریانس نمونه ماکزیمم قرار دارند . شکل ۸-۲ (ب) یک بیضی با فاصلهٔ ثابت متمرکز در $\bar{\mathbf{x}}$ را با $\hat{\lambda}_1 = \hat{\lambda}_2$ نشان می دهد. در این حالت محورهای بیضی (دایره) با چگالی ثابت به طور منحصر به فردی تعیین نمی شوند و می توانند در هر دو جهت متعامد از جمله در جهات محورهای مختصات اولیه قرار گیرند .

به طور مشابه مؤلفه های اصلی نمونه می تواند در هر دو جهت متعامد از جمله جهات محورهای مختصات اولیه واقع شود . وقتی مسیرهای با فاصلهٔ ثابت تقریباً مدورند یا معادل با آن وقتی مقادیر ویژه S تقریباً مساویند ، تغییرات نمونه در تمام جهات همگن است . در این صورت داده ها را نمی توان

در کمتر از p جهت نشان داد .



(a) $\hat{\lambda}_1 > \hat{\lambda}_2$ (b) $\hat{\lambda}_1 \doteq \hat{\lambda}_2$

شکل ۸-۲ مؤلفه‌های اصلی نمونه و بیضی‌های با فاصلهٔ ثابت

اگر $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ را بتوان به صورت نمونه ای از یک جامعهٔ نرمال تلقی کرد ، در آن صورت مؤلفه های اصلی نمونه ، $\hat{y}_i = \hat{\mathbf{e}}'_i(\mathbf{x}-\bar{\mathbf{x}})$ مصادیق مؤلفه های اصلی جامعه ، $Y_i = \mathbf{e}'_i(\mathbf{X}-\boldsymbol{\mu})$ هستند که دارای توزیع $N_p(\mathbf{0}, \mathbf{\Lambda})$ هستند . ماتریس قطری Λ دارای درایه های $\lambda_1, \lambda_2, \ldots, \lambda_p$ است و $(\lambda_i, \mathbf{e}_i)$ زوجهای مقدار ویژه ـ بردار ویژه Σ هستند . همچنین بیضویهای با فاصلهٔ ثابت (۸-۲۴) برآوردهای بیضویهای با چگالی ثابت $(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$ هستند . فرض نرمال برای روشهای استنباطی مورد بحث در (۸-۵) مفید است ، ولی برای تعمیم خواص مؤلفه های اصلی نمونهٔ خلاصه شده در (۸-۲۰) لازم نیستند .

به طور کلی مؤلفه های اصلی نمونه نسبت به تغییرات در مقیاس (تمرین ۸ . ۲ را ملاحظه نمایید) پایا نیستند . به طوری که در عملکرد مؤلفه های جامعه متذکر شدیم ، متغیرهایی که با مقیاسهای مختلف یا مقیاس مشترك ولی در فاصله های با تغییر زیاد اندازه گیری می شوند را اغلب استاندارد می کنیم . استاندارد کردن را برای نمونه با ساختن

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \dfrac{x_{1j}-\bar{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{2j}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \dfrac{x_{pj}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, \ldots, n \tag{۸-۲۵}$$

انجام می دهیم

ماتریس داده های $p \times n$ مشاهدات استاندارد شده

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n] = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pn} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{12} - \bar{x}_1}{\sqrt{s_{11}}} & \cdots & \dfrac{x_{1n} - \bar{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{21} - \bar{x}_2}{\sqrt{s_{22}}} & \dfrac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{2n} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{x_{p1} - \bar{x}_p}{\sqrt{s_{pp}}} & \dfrac{x_{p2} - \bar{x}_p}{\sqrt{s_{pp}}} & \cdots & \dfrac{x_{pn} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \quad (۲۶\text{-}۸)$$

بردار میانگین نمونهٔ [(۳-۲۴) را ملاحظه کنید] ،

$$\bar{\mathbf{z}} = \frac{1}{n}\mathbf{Z}\mathbf{1} = \frac{1}{n} \begin{bmatrix} \displaystyle\sum_{j=1}^{n} \frac{x_{1j} - \bar{x}_1}{\sqrt{s_{11}}} \\ \displaystyle\sum_{j=1}^{n} \frac{x_{2j} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \displaystyle\sum_{j=1}^{n} \frac{x_{pj} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \mathbf{0} \quad (۲۷\text{-}۸)$$

و ماتریس کوواریانس نمونهٔ [(۳-۲۷) را ملاحظه نمایید] ،

$$\mathbf{S}_z = \frac{1}{n-1}\left(\mathbf{Z} - \frac{1}{n}\mathbf{Z}\mathbf{1}\mathbf{1}'\right)\left(\mathbf{Z} - \frac{1}{n}\mathbf{Z}\mathbf{1}\mathbf{1}'\right)' = \frac{1}{n-1}(\mathbf{Z} - \bar{\mathbf{z}}\mathbf{1}')(\mathbf{Z} - \bar{\mathbf{z}}\mathbf{1}')'$$

$$= \frac{1}{n-1}\mathbf{Z}\mathbf{Z}'$$

$$= \frac{1}{n-1} \begin{bmatrix} \dfrac{(n-1)s_{11}}{s_{11}} & \dfrac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \dfrac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \dfrac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \dfrac{(n-1)s_{22}}{s_{22}} & \cdots & \dfrac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \dfrac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \dfrac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix} = \mathbf{R} \quad (۲۸\text{-}۸)$$

را می دهد . مؤلفه های اصلی نمونه مشاهدات استاندارد شده به وسیلهٔ (۸-۲۰) که در آن ماتریس R جانشین S شده است داده می شود . چون مشاهدات را پیشتر «متمرکز» نموده ایم ، لذا نوشتن مؤلفه ها به صورت (۸-۲۱) لزومی ندارد .

اگر $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ مشاهدات استاندارد شده با ماتریس کوواریانس **R** باشند ، مؤلفهٔ اصلی نمونهٔ iام ، عبارت است از :

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z} = \hat{e}_{1i}z_1 + \hat{e}_{2i}z_2 + \cdots + \hat{e}_{pi}z_p, \qquad i = 1, 2, \ldots, p$$

که در آن $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ زوج مقدار ویژه ـ بردار ویژهٔ i ام **R** با $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$ است . همچنین

$$\text{واریانس نمونه } (\hat{y}_i) = \hat{\lambda}_i, \qquad i = 1, 2, \ldots, p$$

$$\text{کوواریانس نمونه } (\hat{y}_i, \hat{y}_k) = 0, \qquad i \neq k$$

علاوه بر این

$$\text{کل واریانس نمونه (استانداردشده)} = \text{tr}(\mathbf{R}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

و

$$r_{\hat{y}_i, z_k} = \hat{e}_{ki}\sqrt{\hat{\lambda}_i}, \qquad i, k = 1, 2, \ldots, p$$

با استفاده از (۸-۲۹) نسبت کل واریانس نمونهٔ بیان شده با مؤلفهٔ اصلی نمونهٔ i ام ، عبارت است از :

$$\left(\begin{array}{c}\text{سهم واریانس نمونه (استانداردشده)}\\ \text{مربوط به مؤلفهٔ اصلی نمونهٔ}\\ i\text{ام}\end{array}\right) = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, \ldots, p \tag{۸-۳۰}$$

یک قانون سرانگشتی پیشنهاد می کند که فقط آن مؤلفه هایی را که واریانس $\hat{\lambda}_i$ آنها بزرگتر از واحد است یا معادل با آن فقط مؤلفه هایی را که به تنهایی حداقل نسبت $\frac{1}{p}$ واریانس کل را بیان می کنند ، نگاه داریم . با این وجود این قانون را نظریه زیاد تأیید نمی کند و آن را نباید بدون دلیل به کار برد .

مثال ۸-۵

نرخهای هفتگی سوددهی پنج سهم (الاید کمیکال ، دوپونت ، یونیون کارباید ، اکسون و تکزاکو) در بازار بورس نیویورک که برای دورهٔ ژانویهٔ ۱۹۷۵ تا دسامبر ۱۹۷۶ تعیین شده است را ثبت نموده ایم . نرخهای هفتگی سوددهی را به صورت (قیمت سهام این جمعه ـ قیمت سهام جمعه

قبل) / (قیمت سهام جمعهٔ قبل) برای خرد کردن و تقسیم سهام تعدیل می شوند .

داده ها در جدول (۸-۱) مربوط به تمرینها ثبت شده است . مشاهدات درصد هفتهٔ متوالی به نظر می رسد مستقلاً توزیع شده اند ، ولی نرخهای سوددهی در **سرتاسر** سهام همبسته اند ، زیرا انتظار می رود سهام در پاسخگویی به کل شرایط اقتصادی با یکدیگر تغییر کنند .

فرض کنید $x_1, x_2, \ldots, x_5$ به ترتیب نرخهای هفتگی سوددهی مشاهده شدهٔ شرکتهای الاید کیمیکال ، دوپونت ، یونیدن کارباید ، اکسون و تکزاکو باشد . در این صورت

$$\bar{\mathbf{x}}' = [.0054, .0048, .0057, .0063, .0037]$$

و

$$\mathbf{R} = \begin{bmatrix} 1.000 & .577 & .509 & .387 & .462 \\ .577 & 1.000 & .599 & .389 & .322 \\ .509 & .599 & 1.000 & .436 & .426 \\ .387 & .389 & .436 & 1.000 & .523 \\ .462 & .322 & .426 & .523 & 1.000 \end{bmatrix}$$

توجه می کنیم که $\mathbf{R}$ ماتریس کوواریانس مشاهدات استاندارد شده

$$z_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}}, z_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}}, \ldots, z_5 = \frac{x_5 - \bar{x}_5}{\sqrt{s_{55}}}$$

است . مقادیر ویژه و بردارهای ویژهٔ نرمال شدهٔ متناظر $\mathbf{R}$ با رایانه محاسبه شده و در زیر داده می شود :

$$\hat{\lambda}_1 = 2.857, \qquad \hat{\mathbf{e}}_1' = [.464, .457, .470, .421, .421]$$

$$\hat{\lambda}_2 = .809, \qquad \hat{\mathbf{e}}_2' = [.240, .509, .260, -.526, -.582]$$

$$\hat{\lambda}_3 = .540, \qquad \hat{\mathbf{e}}_3' = [-.612, .178, .335, .541, -.435]$$

$$\hat{\lambda}_4 = .452, \qquad \hat{\mathbf{e}}_4' = [.387, .206, -.662, .472, -.382]$$

$$\hat{\lambda}_5 = .343, \qquad \hat{\mathbf{e}}_5' = [-.451, .676, -.400, -.176, .385]$$

با استفاده از متغیرهای استاندارد شده دو مؤلفهٔ اصلی اول نمونه را به دست می آوریم :

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = .464z_1 + .457z_2 + .470z_3 + .421z_4 + .421z_5$$

$$\hat{y}_2 = \hat{\mathbf{e}}_2'\mathbf{z} = .240z_1 + .509z_2 + .260z_3 - .526z_4 - .582z_5$$

این مؤلفه ها که

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right)100\% = \left(\frac{2.857 + .809}{5}\right)100\% = 73\%$$

واریانس نمونهٔ (استاندارد شدهٔ) کل را حاصل می ند ، دارای تعابیر جالبی است . اولین مؤلفه یک مجموع وزن دار شدهٔ با وزنهای (تقریباً) مساوی یا «شاخص» پنج سهم است . این مؤلفه را یک ***مؤلفهٔ بازار-بورس کلی*** یا به طور خلاصه یک ***مؤلفهٔ بازار*** می نامند . (در حقیقت این پنج سهم در متوسط صنعتی دو جونز منظور می شود .

دومین مؤلفه یک مقایسهٔ بین سهام کمیکال (الاید کمیکال ، دوپونت ، و یونیون کارباید) و سهام مربوط به نفت (اکسون و تکزاکو) را می دهد که آن را یک مؤلفهٔ صنعتی می نامند . از این رو می بینیم که بیشتر تغییرات در این سوددهی سهام به خاطر فعالیت بازار و فعالیت ناهمبستهٔ صنعت است . این تغییر رفتار قیمت اوراق توسط کینگ [۱۱] نیز پیشنهاد شده است .

تعبیر بقیهٔ مؤلفه ها آسان نیست و روی هم رفته تغییری را نشان می دهند که احتمالاً مختص هریک از سهام است . به هر تقدیر آنها بخش زیادی از واریانس نمونهٔ کل را بیان نمی کنند .

این مثال موردی را که نگاهداشتن یک مؤلفهٔ ($\hat{y}_2$) مربوط به یک مقدار ویژهٔ کمتر از ۱ معقول به نظر می رسد را نشان می دهد .

مثال ۸-۶

علمای ژنتیک اغلب با خصیصه های ارثی که آنها را در طول عمر حیوانات می توان چندین بار اندازه گیری نمود، سر و کار دارند . وزن $n = 150$ موش ماده را (به گرم) بلافاصله بعد از تولد چهار بچه موش اول آنها به دست آوردیم . بردار میانگین نمونه و ماتریس همبستگی نمونه ، عبارت است از :

$$\bar{\mathbf{x}}' = [39.88, 45.08, 48.11, 49.95]$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & .7501 & .6329 & .6363 \\ .7501 & 1.000 & .6925 & .7386 \\ .6329 & .6925 & 1.000 & .6625 \\ .6363 & .7386 & .6625 & 1.000 \end{bmatrix}$$

مقادیر ویژهٔ این ماتریس ، عبارتند از :

$$\hat{\lambda}_1 = 3.058, \quad \hat{\lambda}_2 = .382, \quad \hat{\lambda}_3 = .342, \quad \text{و} \quad \hat{\lambda}_4 = .217$$

توجه می کنیم که اولین مقدار ویژه تقریباً مساوی $1 + (p-1)\bar{r} = 1 + (4-1)(.6854) = 3.056$ است که $\bar{r}$ میانگین حسابی اعضای غیرقطری $\mathbf{R}$ است . گرچه $\hat{\lambda}_4$ تا اندازه‌ای کمتر از $\hat{\lambda}_2$ و $\hat{\lambda}_3$ است ولی مقادیر ویژهٔ باقی مانده کوچک و تقریباً مساویند . از این رو گواهی داریم که ماتریس همبستگی جامعهٔ متناظر $\boldsymbol{\rho}$ ممکن است از نظر «برابری همبستگیها» شکلی مانند (۱۸-۱۵) را داشته باشد . در مورد این نماد در مثال (۸-۹) بیشتر تفحص می کنیم .

اولین مؤلفهٔ اصلی

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = .49z_1 + .52z_2 + .49z_3 + .50z_4$$

$100\left(\frac{\hat{\lambda}_1}{p}\right)\% = 100\,(3.058\;/4)\,\% = 76\,\%$ واریانس کل را توجیه می کند . گرچه متوسط وزن تولدهای بعدی در طول زمان افزایش می یابد ولی تغییر در وزنها تقریباً به وسیلهٔ مؤلفهٔ اصلی اول با ضرایب (تقریباً) مساوی به خوبی بیان می شود .

تذکر . یک مقدار کم غیرمعمول برای مقدار ویژهٔ آخر ماتریس کوواریانس یا ماتریس همبستگی نمونه می تواند یک وابستگی خطی در مجموعه داده ها که ما متوجه آن نبوده ایم را نشان دهد . اگر این امر روی می دهد یکی (یا بیش از یکی) از متغیرها زائد است و باید حذف شود . وضعیتی را که x_1 ، x_2 و x_3 نمرات زیر آزمون و نمرهٔ کل x_4 مجموع $x_1 + x_2 + x_3$ است را در نظر می گیریم . در این صورت گرچه ترکیب خطی $\mathbf{e}'\mathbf{x} = [1, 1, 1, -1]\,\mathbf{x} = x_1 + x_2 + x_3 - x_4$ همیشه صفر است ولی خطای حاصل از گرد کردن محاسبهٔ مقادیر ویژه ، ممکن است منجر به یک مقدار مخالف صفر کوچک گردد . اگر عبارت خطی که x_4 را به (x_1, x_2, x_3) مربوط می کند ، از اول نادیده گرفته می شد در آن صورت کوچکترین زوج مقدار ویژه ـ بردار ویژه باید سرنخی از وجودش را به ما می داد .

از این رو مقادیر ویژهٔ «بزرگ» و بردارهای ویژه مربوط به آن در تحلیل یک مؤلفهٔ اصلی با اهمیت اند و مقادیر ویژه بسیار نزدیک به صفر را نباید معمولاً نادیده گرفت . بردارهای ویژهٔ مربوط به این مقادیر ویژهٔ اخیر ممکن است به نابستگیهای خطی در مجموعهٔ داده ها که می تواند مسائل تعبیری و محاسبه ای را در یک تحلیل بعدی به وجود آورد ، اشاره کند .

۸-۴ نمودار مؤلفه های اصلی

نمودار مؤلفه های اصلی می تواند مشاهدات مورد شک را آشکار نماید و بررسیهای فرض نرمال را فراهم کند . چون مؤلفه های اصلی ترکیبات خطی متغیرهای اولیه هستند ، لذا تقریباً نرمال بودن آنها نامعقول نیست . وقتی از مؤلفه های اصلی به عنوان داده های ورودی برای تحلیلهای دیگر استفاده می کنیم ، اغلب لازم است توزیع چند مؤلفهٔ اصلی اول را نرمال در نظر بگیریم .

مؤلفه های اصلی آخر می توانند در مورد مشاهدات مشکوك به ما کمک کنند . هر مشاهدهٔ $\mathbf{x}_j$ را می توان به صورت یک ترکیب خطی

$$\begin{aligned}\mathbf{x}_j &= (\mathbf{x}_j'\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{x}_j'\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 + \cdots + (\mathbf{x}_j'\hat{\mathbf{e}}_p)\hat{\mathbf{e}}_p \\ &= \hat{y}_{1j}\hat{\mathbf{e}}_1 + \hat{y}_{2j}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{pj}\hat{\mathbf{e}}_p\end{aligned}$$

روشهای تشخیصی مؤلفه های اصلی را به خوبی می توان برای بررسی فرضهای مربوط به الگوی رگرسیون چندگانه چندمتغیری به کار برد . در حقیقت اگر هر الگو را با هر روش برآورد برازش کنیم در نظر گرفتن

(بردار مقادیر تخمین زده شده (برآوردشده)) − (بردار مشاهده) = بردار باقی مانده

یا

$$\underset{(p\times 1)}{\hat{\boldsymbol{\varepsilon}}_j} = \underset{(p\times 1)}{\mathbf{y}_j} - \underset{(p\times 1)}{\mathbf{z}_j'\hat{\boldsymbol{\beta}}}, \qquad j = 1, 2, \ldots, n \tag{۸-۳۱}$$

برای الگوی خطی چندمتغیری عاقلانه است . مؤلفه های اصلی به دست آمده از ماتریس باقی مانده

$$\frac{\sum_{j=1}^{n} (\hat{\boldsymbol{\varepsilon}}_j - \bar{\hat{\boldsymbol{\varepsilon}}}_j)(\hat{\boldsymbol{\varepsilon}}_j - \bar{\hat{\boldsymbol{\varepsilon}}}_j)'}{n-p} \tag{۸-۳۲}$$

را به همان شکلی که از یک نمونهٔ تصادفی به دست آمدند ، مورد بررسی قرار داد . شما باید آگاه باشید که وابستگیهای خطی در میان باقی مانده های حاصل از یک تحلیل رگرسیون خطی وجود دارند ، لذا مقادیر ویژهٔ آخر در بین خطاهای ناشی از گرد کردن صفر خواهد بود .

۸-۵ استنباطهای مبتنی بر نمونه های بزرگ

دیدیم که مقادیر ویژه و بردارهای ویژهٔ ماتریس کوواریانس (همبستگی) جوهرهٔ یک تحلیل مؤلفه های اصلی است . بردارهای ویژه جهات بیشترین تغییرات و مقادیر ویژه واریانسها را مشخص می کند . هنگامی که چند مقدار ویژهٔ اول خیلی بزرگتر از بقیه هستند ، بیشتر واریانس کل را می توان در کمتر از p بُعد «بیان کرد» .

در عمل تصمیمهایی در مورد کیفیت تقریب مؤلفهٔ اصلی بایستی بر پایهٔ زوجهای مقدار ویژه . بردار ویژهٔ به دست آمده از S یا R ساخته شود . این مقادیر ویژه و بردارهای ویژه به علت تغییرات نمونه گیری با مقادیر ویژه و بردارهای ویژهٔ جامعهٔ مورد بررسی تفاوت می کنند . به دست آوردن توزیعهای نمونه گیری $\hat{\lambda}_i$ و $\hat{\mathbf{e}}_i$ مشکل و خارج از بحث این کتاب است . اگر علاقه مند باشید می توانید برخی از اثباتها را برای توزیعهای نرمال چندمتغیری در [۱] و [۲] و [۴] ببینید . نتایج نمونه های بزرگ مناسب را به آسانی خلاصه می کنیم .

خواص نمونه های بزرگ $\hat{\lambda}_i$ و $\hat{\mathbf{e}}_i$

در نتایج مربوط به فواصل اطمینان نمونه های بزرگ برای $\hat{\lambda}_i$ و $\hat{\mathbf{e}}_i$ که در حال حاضر

در دسترس هستند ، $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ یک نمونهٔ تصادفی از یک جامعهٔ نرمال در نظر گرفته می شود . همچنین باید فرض کنیم که مقادیر ویژهٔ (نامعلوم) $\mathbf{\Sigma}$ متمایز و مثبت اند ، به طوری که $\lambda_1 > \lambda_2 > ... > \lambda_p > 0$. حالتی که تعداد مقادیر ویژه مساوی معلوم اند یک استثناست . معمولاً نتایج مربوط به مقادیر ویژه متمایز به کار برده می شود ، مگر این که یک دلیل قوی وجود داشته باشد که باور کنیم $\mathbf{\Sigma}$ برای حصول مقادیر ویژهٔ مساوی دارای ساختمان ویژه ای است . حتی وقتی از فرض نرمال عدول می شود ، فواصل اطمینانی که به این طریق به دست می آید ، هنوز دلیلی بر عدم قطعیت $\hat{\lambda}_i$ و $\hat{\mathbf{e}}_i$ فراهم می کند .

اندرسن [۲] و گیرشیک [۴] نظریهٔ توزیع نمونه های بزرگ زیر را برای مقادیر ویژهٔ $\hat{\boldsymbol{\lambda}}' = [\hat{\lambda}_1, ..., \hat{\lambda}_p]$ و بردارهای ویژه $\hat{\mathbf{e}}_1, ..., \hat{\mathbf{e}}_p$ ماتریس $\mathbf{S}$ ثابت می کنند .

۱- فرض کنید $\mathbf{\Lambda}$ ماتریس قطری مقادیر ویژهٔ $\lambda_1, ..., \lambda_p$ ، $\mathbf{\Sigma}$ باشد در آن صورت $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ تقریباً $N_p(\mathbf{0}, 2\mathbf{\Lambda}^2)$ است .

۲- فرض کنید

$$\mathbf{E}_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k'$$

در این صورت $\sqrt{n}(\hat{\mathbf{e}}_i - \mathbf{e}_i)$ تقریباً $N_p(\mathbf{0}, \mathbf{E}_i)$ است .

۳- هر $\hat{\lambda}_i$ مستقل از اعضای مربوط به $\hat{\mathbf{e}}_i$ توزیع می شود .

از نتیجهٔ ۱ معلوم می شود که برای n بزرگ ، $\hat{\lambda}_i$ ها مستقلاً توزیع می شوند . علاوه بر این $\hat{\lambda}_i$ دارای یک توزیع تقریبی $N(\lambda_i, \frac{2\lambda_i^2}{n})$ است . با استفاده از این توزیع نرمال ، داریم : $P[|\hat{\lambda}_i - \lambda_i| \le z(\alpha/2)\lambda_i\sqrt{2/n}] = 1 - \alpha$ به این ترتیب یک فاصلهٔ اطمینان $100(1-\alpha)\%$ نمونهٔ بزرگ برای λ_i ، عبارت است از :

$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \le \lambda_i \le \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})} \tag{۸-۳۳}$$

که در آن $z(\alpha/2)$ صدك $100(\alpha/2)$ ام یک توزیع نرمال استاندارد است . فواصل $100(1-\alpha)\%$ هم زمان از نوع بونفرونی برای m تا λ_i با جایگزینی $z(\alpha/2)$ با $z(\alpha/2m)$ به دست می آید (بخش ۵-۴ را ملاحظه نمایید) . از نتیجهٔ ۲ معلوم می شود که برای نمونه های بزرگ $\hat{\mathbf{e}}_i$ ها حول $\mathbf{e}_i$ های مربوطه به طور نرمال توزیع می شوند . اعضای هر $\hat{\mathbf{e}}_i$ همبسته بوده و همبستگی تا حد زیادی به جدایی مقادیر ویژهٔ $\lambda_1, \lambda_2, ..., \lambda_p$ (که نامعلوم اند) و حجم نمونهٔ n بستگی دارد . خطاهای معیار تقریبی ضرایب $\hat{e}_{ki}$

با اعضای قطری $(1/n)\hat{\mathbf{E}}_i$ داده می شود که $\hat{\mathbf{E}}_i$ با قرار دادن $\hat{\lambda}_i$ ها به جای λ_i ها از $\mathbf{E}_i$ به دست می آید .

مثال ۸-۸

با استفاده از داده های قیمت سهام در جدول (۸-۱) یک فاصلهٔ اطمینان 95% برای λ_1 یعنی واریانس اولین مؤلفه اصلی جامعه ، به دست می آوریم .

فرض می کنیم نرخهای سوددهی سهام استخراجهای مستقل از یک $N_5(\mu, \Sigma)$ که Σ معین و مثبت با مقادیر ویژهٔ متمایز $\lambda_1 > \lambda_2 > \ldots > \lambda_5 > 0$ است را نشان دهد . چون $n = 100$ بزرگ است ، می توانیم از (۸-۳۳) با $i = 1$ برای ساختن یک فاصلهٔ اطمینان 95% برای λ_1 استفاده کنیم از تمرین ۸ . ۱۰ داریم $\hat{\lambda}_1 = .0036$ و علاوه بر این $z(.025) = 1.96$. بنابراین با اطمینان 95% ، داریم :

$$\frac{.0036}{(1 + 1.96\sqrt{\frac{2}{100}})} \le \lambda_1 \le \frac{.0036}{(1 - 1.96\sqrt{\frac{2}{100}})} \quad \text{یا} \quad .0028 \le \lambda_1 \le .0050$$

هرگاه یک مقدار ویژه مانند ۱۰۰ یا حتی ۱۰۰۰ بزرگ باشد فواصلی که به وسیلهٔ (۸-۳۳) تولید می شوند برای سطوح اطمینان معقولی می توانند کاملاً عریض باشند ولو این که n نسبتاً بزرگ باشد . به طور کلی فاصلهٔ اطمینان به همان نرخی که $\hat{\lambda}_i$ بزرگتر می شود عریض تر می گردد . بنابراین باید بر پایهٔ یک امتحان $\hat{\lambda}_i$ ها در حذف یا نگهداری مؤلفه های اصلی دقت کنیم .

آزمون مربوط به ساختار همبستگیهای مساوی

ساختار ویژه همبستگی $\operatorname{Cov}(X_i, X_k) = \sqrt{\sigma_{ii}\sigma_{kk}}\,\rho$ یا $\operatorname{Corr}(X_i, X_k) = \rho$ برای هر $i \neq k$ ساختار مهمی است که مقادیر ویژهٔ $\boldsymbol{\Sigma}$ متمایز نیست و نتایج قبلی کاربرد نداشته باشند .

برای آزمون این ساختار ، فرض کنید :

$$H_0: \boldsymbol{\rho} = \underset{(p\times p)}{\boldsymbol{\rho}_0} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

و

$$H_1: \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

یک آزمون H_0 در مقابل H_1 می تواند بر پایهٔ یک آمارهٔ نسبت درست نمایی قرار داشته باشد ، ولی لاولی [۱۲] نشان می دهد که می توان یک روش آزمون معادلی را از اعضای غیرقطری $\mathbf{R}$ ساخت .

در روش لاولی کمیتهای زیر لازم است :

$$\bar{r}_k = \frac{1}{p-1}\sum_{\substack{i=1\\ i\neq k}}^{p} r_{ik} \qquad k = 1, 2, \ldots, p; \qquad \bar{r} = \frac{2}{p(p-1)}\sum\sum_{i<k} r_{ik} \tag{۳۴-۸}$$

$$\hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}$$

واضح است که $\bar{r}_k$ متوسط اعضای غیرقطری در ستون (یا سطر) k ام $\mathbf{R}$ و $\bar{r}$ متوسط تمام اعضای غیرقطری است .

آزمون تقریبی نمونه های بزرگ در سطح α دارای شکل زیر است : H_0 را به نفع H_1 رد می کنیم هرگاه

$$T = \frac{(n-1)}{(1-\bar{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2\right] > \chi^2_{(p+1)(p-2)/2}(\alpha) \tag{۳۵-۸}$$

که در آن $\chi^2_{(p+1)(p-2)/2}(\alpha)$ صدك (100α) ام بالایی یک توزیع کی دو با $\frac{(p+1)(p-2)}{2}$ درجهٔ آزادی است .

مثال ۸-۹

ماتریس همبستگی نمونه ساخته شده از وزنهای تولد بعدی موشهای ماده در مثال (۸-۶) به شرح زیر است :

$$\mathbf{R} = \begin{bmatrix} 1.0 & .7501 & .6329 & .6363 \\ .7501 & 1.0 & .6925 & .7386 \\ .6329 & .6925 & 1.0 & .6625 \\ .6363 & .7386 & .6625 & 1.0 \end{bmatrix}$$

ما از این ماتریس همبستگی برای تشریح آزمون نمونهٔ بزرگ در (۸-۳۵) استفاده خواهیم کرد .

در این جا $p = 4$ بوده و قرار می دهیم :

$$H_0: \quad \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

$$H_1: \quad \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

با استفاده از (۸-۳۴) و (۸-۳۵) ، داریم :

$$\bar{r}_1 = \tfrac{1}{3}(.7501 + .6329 + .6363) = .6731, \qquad \bar{r}_2 = .7271,$$
$$\bar{r}_3 = .6626, \qquad \bar{r}_4 = .6791$$
$$\bar{r} = \frac{2}{4(3)}(.7501 + .6329 + .6363 + .6925 + .7386 + .6625) = .6855$$
$$\sum\sum_{i<k} (r_{ik} - \bar{r})^2 = (.7501 - .6855)^2 + (.6329 - .6855)^2 + \cdots + (.6625 - .6855)^2 = .01277$$
$$\sum_{k=1}^{4} (\bar{r}_k - \bar{r})^2 = (.6731 - .6855)^2 + \cdots + (.6791 - .6855)^2 = .00245$$
$$\hat{\gamma} = \frac{(4-1)^2[1-(1-.6855)^2]}{4-(4-2)(1-.6855)^2} = 2.1329$$

و

$$T = \frac{(150-1)}{(1-.6855)^2}[.01277 - (2.1329)(.00245)] = 11.4$$

چون $(p+1)(p-2)/2 = 5(2)/2 = 5$ لذا مقدار بحرانی 5% برای آزمون در (۸-۳۵) $\chi^2_5(.05) = 11.07$ است. مقدار آمارهٔ آزمون ما تقریباً مساوی نقطهٔ بحرانی 5% نمونه بزرگ است، لذا شواهد علیه H_0 (همبستگیهای مساوی) قوی بوده ولی بیش از حد اندازه نیست. چنان که در مثال (۸-۶) دیدیم، کوچکترین مقادیر ویژه $\hat{\lambda}_2$، $\hat{\lambda}_3$ و $\hat{\lambda}_4$ قدری با هم تفاوت دارند و $\hat{\lambda}_4$ تا اندازه‌ای کمتر از دوتای دیگر است. در نتیجه با حجم نمونهٔ بزرگی که در این مسأله است تفاوتهای کم از ساختار همبستگیهای مساوی یک معنی داری آماری را نشان می دهند.

تمرینها

۱.۸ مؤلفه های اصلی Y_1 و Y_2 جامعه را برای ماتریس کوواریانس

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

SAS

$\frac{2}{\sqrt{5}\sqrt{2}}$

تعیین کنید همچنین نسبت واریانس کل جامعه بیان شده با اولین مؤلفهٔ اصلی را محاسبه کنید .

$\begin{bmatrix} 1 & 0.6 \approx .4455332 \\ & 1 \end{bmatrix}$

۲.۸ ماتریس کوواریانس در تمرین (۱.۸) را به ماتریس همبستگی $\boldsymbol{\rho}$ تبدیل کنید .

(الف) مؤلفه های اصلی Y_1 و Y_2 را از $\boldsymbol{\rho}$ تعیین کرده و نسبت واریانس کل جامعه بیان شده با Y_1 را محاسبه کنید .

(ب) مؤلفه ها را با مؤلفه های اصلی به دست آمده در تمرین (۱.۸) مقایسه کنید . آیا این دو یکسانند ؟ آیا باید یکسان باشند ؟

(ج) همبستگیهای ρ_{Y_1, Z_1} ، ρ_{Y_1, Z_2} و ρ_{Y_2, Z_1} را محاسبه کنید .

۳.۸ فرض کنید :

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

مؤلفه های اصلی Y_1 و Y_2 و Y_3 را تعیین کنید . نظر شما در مورد بردارهای ویژه (و مؤلفه های اصلی) مربوط به مقادیر ویژه ای که متمایز نیستند چیست ؟

۴.۸ مؤلفه های اصلی ونسبت واریانس کل جامعه بیان شده توسط هریک را وقتی ماتریس کوواریانس

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix}, \qquad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$$

پیدا کنید .

۵.۸ (الف) مقادیر ویژهٔ ماتریس همبستگی زیر را پیدا کنید ،

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

آیا نتایج شما با (۸-۱۶) و (۸-۱۷) سازگار است ؟

(ب) زوجهای مقدار ویژه ـ بردار ویژه را برای ماتریس $p \times p$ ، $\boldsymbol{\rho}$ داده شده در (۸-۱۵) به دست آورید .

۶ . ۸ داده های مربوط به دارایی = x_1 و درآمد خالص = x_2 برای ده تا از بزرگترین شرکتهای صنعتی ایالات متحده در تمرین (۱ . ۴) فصل ۱ ثبت شده است .

از مثال (۴-۱۱) ، داریم :

$$\bar{\mathbf{x}} = \begin{bmatrix} 19.32 \\ 1.51 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} 70.41 & 5.87 \\ 5.87 & .97 \end{bmatrix}$$

(الف) مؤلفه های اصلی نمونه و واریانس آنها را برای این داده ها تعیین کنید (شما ممکن است برای به دست آوردن مقادیر ویژهٔ S یک فرمول درجهٔ دوم لازم داشته باشید) .

(ب) نسبت واریانس کل نمونه بیان شده با $\hat{y}_1$ را پیدا کنید .

(ج) بیضی با چگالی ثابت $(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) = 1.4$ را رسم کرده و مؤلفه های اصلی $\hat{y}_1$ و $\hat{y}_2$ را روی نمودار نشان دهید .

(د) ضرایب همبستگی $r_{\hat{y}_1, x_k}$ ، $k = 1, 2$ را محاسبه کنید . چه تعبیری از اولین مؤلفهٔ اصلی ارائه می کنید ؟

۷ . ۸ ماتریس کوواریانس S را در تمرین (۸ . ۶) به یک ماتریس همبستگی نمونهٔ **R** تبدیل کنید .

(الف) مؤلفه های اصلی نمونهٔ $\hat{y}_1$ و $\hat{y}_2$ و واریانسهایشان را پیدا کنید .

(ب) نسبت واریانس کل نمونه بیان شده با $\hat{y}_1$ را محاسبه کنید .

(ج) ضرایب همبستگی $r_{\hat{y}_1, x_k}$ ، $k = 1, 2$ را محاسبه نموده و $\hat{y}_1$ را تعبیر کنید .

(د) مؤلفه های به دست آمده در بخش الف را با مؤلفه های به دست آمده در تمرین (۸ . ۶) الف مقایسه کنید . داده های اولیه نمایش داده شده در تمرین (۱ . ۴) معلوم اند، آیا احساس می کنید بهتر است مؤلفه های اصلی را از ماتریس کوواریانس نمونه یا ماتریس همبستگی نمونه تعیین کنیم ؟ بیان کنید .

۸ . ۸ با استفاده از نتایج مثال (۸-۵)

(الف) همبستگیهای $r_{\hat{y}_i, z_k}$ ، را برای $i = 1, 2$ ، $k = 1, 2, \ldots, 5$ محاسبه کنید . آیا این همبستگیها تعابیر ارائه شده برای دو مؤلفهٔ اول را تقویت می کند ؟

(ب) فرض

$$H_0: \quad \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

در مقابل