How to check if samples are Gaussian

Guglielmo Padula

March 13, 2024

Let's suppose that we have a dataset $\mathbf{X} \in \mathbb{R}^{N_{samples} \times N_{dim}}$ of rank $N_{latent} \leq \min\{N_{samples}, N_{dim}\}$ and $\mathbf{X} \sim \mathcal{X}$. We would check if $\mathcal{X}$ is a (degenerate if $N_{dim} > N_{latent}$) normal distribution. Let $\bar{\mathbf{X}} = \frac{\mathbf{X} - mean(\mathbf{X}, axis=0)}{\sqrt{N_{samples}-1}}$ and let $\bar{\mathbf{X}}^T = \mathbf{USV}$ an SVD of $\bar{\mathbf{X}}^T$ with $\mathbf{U} \in \mathbb{R}^{N_{dim} \times N_{latent}}$, $\mathbf{S} \in \mathbb{R}^{N_{latent} \times N_{latent}}$, $\mathbf{V} \in \mathbb{R}^{N_{latent} \times N_{samples}}$ Let $\bar{\mathbf{B}} = \mathbf{US} \in \mathbb{R}^{N_{dim} \times N_{latent}}$.
Note that $Cov(\mathcal{X}) \approx \bar{\mathbf{X}}^{\mathbf{T}}\bar{\mathbf{X}} = \bar{\mathbf{B}}\bar{\mathbf{B}}^{\mathbf{T}}$.
Let $\mu = Mean(\mathcal{X})$.
Let $\bar{\mathbf{B}}^{+} = \mathbf{S}^{-1}\mathbf{U}^T \in \mathbb{R}^{N_{latent} \times N_{dim}}$. Let $\mathcal{Z} = \bar{\mathbf{B}}^{+}(\mathcal{X} - \mu)$. Because of the properties of the covariance we have that $Cov(\mathcal{Z}) = \bar{\mathbf{B}}^{+}Cov(X)\bar{\mathbf{B}}^{T+} = \mathbf{I}_{N_{latent}}$ Also it holds $\bar{\mathbf{B}}\mathcal{Z} + \mu = \bar{\mathbf{B}}(\bar{\mathbf{B}}^{+}\mathcal{X} - \mu) + \mu = \mathbf{I}_{N_{dim}}\mathcal{X} = \mathcal{X}$. Because of the properties of the multivariate normal distributions, we have that $\mathcal{X}$ is normal if and only if $\mathcal{Z}$ is normal. So we can check the normality of $\mathcal{X}$ by checking the samples $\mathbf{Z} = (\mathbf{X} - \mu)\mathbf{B}^{T+}$. This is easy is we use the empirical cdf confidence interval

$$F_n(x) - \varepsilon \leq F(x) \leq F_n(x) + \varepsilon \text{ where } \varepsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}}$$

of the Dvoretzky–Kiefer–Wolfowitz inequality.