A   Generalised inverse Gaussian distribution

The generalised inverse Gaussian distribution is defined as follows:

$$x \sim \mathcal{N}^{-1}(\omega, \chi, \phi) = \frac{\chi^{-\omega}(\sqrt{\chi\phi})^{\omega}}{2K_{\omega}(\sqrt{\chi\phi})} x^{\omega-1} e^{-\frac{1}{2}(\chi x^{-1} + \phi x)}, \tag{1}$$

where $x > 0$ and $K_{\omega}(\cdot)$ is the modified Bessel function of the second kind with index $\omega \in \mathbb{R}$. Depending on the value taken by ω, we have the following constraints on χ and ϕ:

$$\begin{cases} \omega > 0: & \chi \geqslant 0, \ \phi > 0, \\ \omega = 0: & \chi > 0, \ \phi > 0, \\ \omega < 0: & \chi > 0, \ \phi \geqslant 0. \end{cases}$$

The following expectations are useful [**?**]:

$$\langle x \rangle = \sqrt{\frac{\chi}{\phi}} R_{\omega}(\sqrt{\chi\phi}), \quad \langle x^{-1} \rangle = \sqrt{\frac{\phi}{\chi}} R_{-\omega}(\sqrt{\chi\phi}), \quad \langle \ln x \rangle = \ln\sqrt{\frac{\chi}{\phi}} + \frac{d \ln K_{\omega}(\sqrt{\chi\phi})}{d\omega}, \quad (2)$$

where $R_{\omega}(\cdot) \equiv K_{\omega+1}(\cdot)/K_{\omega}(\cdot)$

When $\chi = 0$ and $\omega > 0$, the generalised inverse Gaussian distribution reduces to the Gamma distribution $x \sim \mathcal{G}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$, where $a, b > 0$ and $\Gamma(\cdot)$ is the (complete) gamma function. The expectations (2) take the following simplified forms:

$$\langle x \rangle = \frac{a}{b}, \qquad \langle x^{-1} \rangle = \begin{cases} \frac{b}{a-1} & a \geqslant 1 \\ \infty & a < 1 \end{cases}, \qquad \langle \ln x \rangle = \psi(a) - \ln b, \tag{3}$$

where $\psi(\cdot) = \ln \Gamma(\cdot)'$ is the digamma function.

When $\phi = 0$ and $\omega < 0$, the generalised inverse Gaussian distribution reduces to the inverse Gamma distribution $x \sim \mathcal{IG}(a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-\frac{b}{x}}$, where $a > 0$ and $b > 0$. The expectations (2) take the following simplfied forms:

$$\langle x \rangle = \begin{cases} \frac{b}{a-1} & a \geqslant 1 \\ \infty & a < 1 \end{cases}, \qquad \langle x^{-1} \rangle = \frac{a}{b}, \qquad \langle \ln x \rangle = \ln b - \psi(a). \tag{4}$$

B   Truncated Gaussian density

The (positive/negative) truncated Gaussian density is defined as $\mathcal{N}_{\pm}(\mu, \sigma^2) = \Phi(\pm\mu/\sigma)^{-1} \mathcal{N}(\mu, \sigma^2)$, where $\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(0, 1) dz$ is the cumulative density of the unit Gaussian.

Let $x_{\pm} \sim \mathcal{N}_{\pm}(\mu, \sigma^2)$. The mean and variance are given by

$$\langle x_{\pm} \rangle = \mu \pm \sigma^2 \mathcal{N}_{\pm}(0|\mu, \sigma^2), \tag{5}$$

$$\langle (x_{\pm} - \langle x_{\pm} \rangle)^2 \rangle = \sigma^2 \mp \sigma^2 \mu \mathcal{N}_{\pm}(0|\mu, \sigma^2) - \sigma^4 \mathcal{N}_{\pm}(0|\mu, \sigma^2)^2. \tag{6}$$

C   Posterior parameters

Let $\mathbf{X} \in \mathbb{R}^{D \times N}$ and $\mathbf{Y} \in \mathbb{R}^{P \times N}$. The parameters of the matrix-variate posteriors in (**??**) are given by

$$\mathbf{M}_W = \left(\tau^{-1} \langle \mathbf{V} \rangle \langle \mathbf{Z} \rangle \langle \mathbf{\Omega}^{-1} \rangle \langle \mathbf{\Gamma} \rangle + \sigma^{-2} \mathbf{Y} \mathbf{X}^{\top} \right) \mathbf{\Omega}_W, \qquad \mathbf{S}_W = \mathbf{I}_P \tag{7}$$

$$\mathbf{\Omega}_W = \left(\tau^{-1} \langle \mathbf{\Omega}^{-1} \rangle \langle \mathbf{\Gamma} \rangle + \sigma^{-2} \mathbf{X} \mathbf{X}^{\top} \right)^{-1},$$

$$\mathbf{M}_{Z_i} = \tau^{-1} \mathbf{S}_{Z_i} \langle \mathbf{V}^{\top} \rangle \langle \mathbf{W}_i \rangle, \qquad\qquad \mathbf{S}_{Z_i} = \left(\tau^{-1} \langle \mathbf{V}^{\top} \mathbf{V} \rangle + \mathbf{I}_K \right)^{-1}, \tag{8}$$

$$\mathbf{\Omega}_{Z_i} = \langle \gamma_i \rangle^{-1} \langle \mathbf{\Omega}_i^{-1} \rangle^{-1},$$

$$\mathbf{M}_V = \langle \mathbf{W} \rangle \langle \mathbf{\Omega}^{-1} \rangle \langle \mathbf{\Gamma} \rangle \langle \mathbf{Z}^{\top} \rangle \mathbf{\Omega}_V, \qquad\qquad \mathbf{S}_V = \tau \mathbf{I}_P, \tag{9}$$

$$\mathbf{\Omega}_V = \left(\sum_i \langle \gamma_i \mathbf{Z}_i \mathbf{\Omega}_i^{-1} \mathbf{Z}_i^\top \rangle + \mathbf{I}_K\right)^{-1}.$$

where $\langle \mathbf{\Omega}_i^{-1} \rangle = (D_i + \upsilon_i - 1)\mathbf{\Lambda}_i^{-1}$.

The posterior parameters of the inverse Wishart are given by $\upsilon_i = \upsilon + P + K$ and $\mathbf{\Lambda}_i = \tau^{-1}\langle \gamma_i(\mathbf{W}_i - \mathbf{VZ}_i)^\top(\mathbf{W}_i - \mathbf{VZ}_i)\rangle + \langle \gamma_i \mathbf{Z}_i^\top \mathbf{Z}_i\rangle + \lambda \mathbf{I}_{D_i}$.

Finally, tthe posterior parameters of the generalised inversed Gaussian are given by $\omega_i = \omega + \frac{(P+K)D_i}{2}$, $\chi_i = \chi$ and $\phi_i = \phi + \tau^{-1}\mathrm{tr}\langle(\mathbf{W}_i - \mathbf{VZ}_i)\mathbf{\Omega}_i^{-1}(\mathbf{W}_i - \mathbf{VZ}_i)^\top\rangle + \mathrm{tr}\langle \mathbf{Z}_i \mathbf{\Omega}_i^{-1} \mathbf{Z}_i^\top \rangle$.

D Variational bound

The variational free energy is an upper bound to the marginal log-likelihood $\ln p(\mathcal{D}|\vartheta)$ [?]. The variational bound for sparse multiple classification model defined in (??) and (??) is given by

$$
\begin{aligned}
\mathcal{F}_q(\mathcal{D}, \mathcal{Z}, \vartheta) = & - \langle \ln p(\mathbf{T}|\mathbf{Y})\rangle - \langle \ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W})\rangle - \langle \ln p(\mathbf{W}|\mathbf{Z}, \mathbf{\Omega}, \mathbf{\Gamma})\rangle - \langle \ln p(\mathbf{Z}|\mathbf{\Omega}, \mathbf{\Gamma})\rangle \\
& - \sum_i \langle \ln p(\mathbf{\Omega}_i)\rangle - \langle \ln p(\mathbf{V})\rangle - \sum_i \langle \ln p(\gamma_i)\rangle + \langle \ln q(\mathbf{Y})\rangle + \langle \ln q(\mathbf{W})\rangle \\
& + \langle \ln q(\mathbf{Z})\rangle + \sum_i \langle \ln q(\mathbf{\Omega}_i)\rangle + \langle \ln q(\mathbf{V})\rangle + \sum_i \langle \ln q(\gamma_i)\rangle,
\end{aligned}
$$

where

$$\langle \ln p(\mathbf{T}|\mathbf{Y})\rangle = -\sum_n \sum_p \langle \ln I(t_{np}y_{np})\rangle$$

$$\langle \ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W})\rangle = -\frac{NP}{2}\ln 2\pi - \frac{NP}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\mathrm{tr}\langle(\mathbf{Y} - \mathbf{WX})^\top(\mathbf{Y} - \mathbf{WX})\rangle$$

$$
\begin{aligned}
\langle \ln p(\mathbf{W}|\mathbf{Z}, \mathbf{\Omega}, \mathbf{\Gamma})\rangle = & -\frac{DP}{2}\ln 2\pi + \frac{P}{2}\langle \ln|\mathbf{\Gamma}|\rangle + \frac{P}{2}\langle \ln|\mathbf{\Omega}^{-1}|\rangle - \frac{DP}{2}\ln \tau \\
& - \frac{1}{2\tau}\mathrm{tr}\langle \mathbf{\Gamma}\mathbf{\Omega}^{-1}(\mathbf{W} - \mathbf{VZ})^\top(\mathbf{W} - \mathbf{VZ})\rangle
\end{aligned}
$$

$$\langle \ln p(\mathbf{Z}|\mathbf{\Omega}, \mathbf{\Gamma})\rangle = -\frac{DK}{2}\ln 2\pi + \frac{K}{2}\langle \ln|\mathbf{\Gamma}|\rangle + \frac{K}{2}\langle \ln|\mathbf{\Omega}^{-1}|\rangle - \frac{1}{2}\mathrm{tr}\langle \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{Z}^\top\mathbf{Z}\rangle$$

$$
\begin{aligned}
\sum_i \langle \ln p(\mathbf{\Omega}_i)\rangle = & -\frac{\sum_i D_i(\upsilon + D_i - 1)}{2}\ln 2 - \frac{\sum_i D_i(D_i - 1)}{4}\ln \pi - \sum_i \sum_{j=1}^{D_i}\ln \Gamma\left(\frac{\upsilon + D_i - j}{2}\right) \\
& + \frac{\sum_i D_i(\upsilon + D_i - 1)}{2}\ln \lambda + \sum_i \frac{\upsilon + 2D_i}{2}\langle \ln|\mathbf{\Omega}_i^{-1}|\rangle - \frac{\lambda}{2}\sum_i \mathrm{tr}\langle \mathbf{\Omega}_i^{-1}\rangle
\end{aligned}
$$

$$\langle \ln p(\mathbf{V})\rangle = -\frac{PK}{2}\ln 2\pi - \frac{PK}{2}\ln \tau - \frac{1}{2\tau}\mathrm{tr}\langle \mathbf{V}^\top\mathbf{V}\rangle$$

$$\sum_i \langle \ln p(\gamma_i)\rangle = Q\omega \ln \sqrt{\frac{\phi}{\chi}} - Q\ln 2K_\omega\left(\sqrt{\chi\phi}\right) + (\omega - 1)\sum_i \langle \ln \gamma_i\rangle - \frac{1}{2}\sum_i\left(\chi\langle \gamma_i^{-1}\rangle + \phi\langle \gamma_i\rangle\right)$$

$$\langle \ln q(\mathbf{Y})\rangle = -\sum_n \sum_p \ln \Phi\left(\frac{t_{np}\nu_{np}}{\sigma}\right) - \frac{NP}{2}\ln 2\pi - \frac{NP}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\mathrm{tr}\langle(\mathbf{Y} - \boldsymbol{\nu})^\top(\mathbf{Y} - \boldsymbol{\nu})\rangle$$

$$\langle \ln q(\mathbf{W})\rangle = -\frac{DP}{2}\ln 2\pi e - \frac{P}{2}\ln|\mathbf{\Omega}_W| - \frac{D}{2}\ln|\mathbf{S}_W|,$$

$$\langle \ln q(\mathbf{Z})\rangle = -\frac{DK}{2}\ln 2\pi e - \frac{K}{2}\ln|\mathbf{\Omega}_Z| - \frac{D}{2}\ln|\mathbf{S}_Z|,$$

$$
\begin{aligned}
\sum_i \langle \ln q(\mathbf{\Omega}_i)\rangle = & -\frac{\sum_i D_i(\upsilon_i + D_i - 1)}{2}\ln 2 - \frac{\sum_i D_i(D_i - 1)}{4}\ln \pi - \sum_i \sum_{j=1}^{D_i}\ln \Gamma\left(\frac{\upsilon_i + D_i - j}{2}\right) \\
& + \sum_i \frac{\upsilon_i + D_i - 1}{2}\ln|\mathbf{\Lambda}_i| + \sum_i \frac{\upsilon_i + 2D_i}{2}\langle \ln|\mathbf{\Omega}_i^{-1}|\rangle - \frac{1}{2}\sum_i \mathrm{tr}\langle \mathbf{\Lambda}_i\mathbf{\Omega}_i^{-1}\rangle
\end{aligned}
$$

$$\langle \ln q(\mathbf{V})\rangle = -\frac{PK}{2}\ln 2\pi e - \frac{P}{2}\ln|\mathbf{\Omega}_V| - \frac{K}{2}\ln|\mathbf{S}_V|,$$

$$\sum_i \langle \ln q(\gamma_i)\rangle = \sum_i \omega_i \ln \sqrt{\frac{\phi_i}{\chi_i}} - \sum_i \ln 2K_{\omega_i}\left(\sqrt{\chi_i\phi_i}\right) + \sum_i(\omega_i - 1)\langle \ln \gamma_i\rangle$$

$$- \frac{1}{2} \sum_i \left(\chi_i \langle \gamma_i^{-1} \rangle + \phi_i \langle \gamma_i \rangle \right).$$

The expectation $\langle \ln |\mathbf{\Omega}_i^{-1}| \rangle$ can be computed as $D_i \ln 2 + \ln |\mathbf{\Lambda}_i| + \sum_j \psi \left(\frac{D_i + v_i - j}{2} \right)$.

In the case of sparse multiple regression model the bound takes a very similar, but simpler form
$\mathcal{F}_q(\mathcal{D}, \mathcal{Z}, \vartheta) = -\langle \ln p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) \rangle - \langle \ln p(\mathbf{W}|\mathbf{Z}, \mathbf{\Omega}, \mathbf{\Gamma}) \rangle - \langle \ln p(\mathbf{Z}|\mathbf{\Omega}, \mathbf{\Gamma}) \rangle - \sum_i \langle \ln p(\mathbf{\Omega}_i) \rangle - \langle \ln p(\mathbf{V}) \rangle - \sum_i \langle \ln p(\gamma_i) \rangle + \langle \ln q(\mathbf{W}) \rangle + \langle \ln q(\mathbf{Z}) \rangle + \sum_i \langle \ln q(\mathbf{\Omega}_i) \rangle + \langle \ln q(\mathbf{V}) \rangle + \sum_i \langle \ln q(\gamma_i) \rangle.$