Acelerando o processo de estimação nas distribuições misturas de escala de normais multivariadas

Autor: Graciliano Márcio Santos Louredo

Instituição: Universidade Federal de Juiz de Fora

Em diversas áreas do conhecimento, ainda é bastante comum recorrer à distribuição normal (gaussiana) para modelar dados. Esse fato decorre em especial da facilidade de estimação dos seus parâmetros, inclusive pelo uso do tradicional método dos mínimos quadrados.

Contudo, a frequência de dados reais com comportamento assimétrico e forte presença de *outliers* torna inviável a modelagem estatística por meio da família locação-escala da distribuição normal. Esse problema motivou o estudo de distribuições assimétricas e, independentemente, das chamadas *distribuições misturas de escala*.

Em particular, destacamos a distribuição normal assimétrica multivariada definida em [1] e as distribuições misturas de escala da família normal propostas em [5], que generalizam a distribuição normal. Na sequência apresentamos a definição formal de cada uma.

Definição 0.1. Diz-se que um vetor aleatório p-dimensional $\mathbf{Y}$ segue uma *distribuição normal assimétrica p-variada* e escrevemos $\mathbf{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ quando sua f.d.p. é dada por

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})),\ \ \mathbf{y} \in \mathbb{R}^p. \tag{1}$$

Acima, denotamos por $\phi_p(\ .\ ; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ a f.d.p. da distribuição normal p-variada com vetor de médias $\boldsymbol{\mu}$ (parâmetro de locação) e matriz de covariâncias $\boldsymbol{\Sigma}$ positiva definida (parâmetro de escala) e por $\Phi_1(\ .\ )$ a f.d.a. da distribuição normal padrão univariada, em cujo argumento figura o vetor $\boldsymbol{\lambda}$ como o parâmetro de assimetria. Observe que a distribuição normal p-variada é obtida fazendo $\boldsymbol{\lambda} = \mathbf{0}$.

Definição 0.2. Diz-se que um vetor aleatório p-dimensional $\mathbf{Y}$ segue uma *distribuição mistura de escala da família normal p-variada* com vetor de locação $\boldsymbol{\mu}$ e matriz de escala $\boldsymbol{\Sigma}$ quando sua fdp é dada por

$$f_0(\mathbf{y}) = \int_0^{+\infty} \phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{u}\right) h(u; \boldsymbol{\tau})du,\ \ \mathbf{y} \in \mathbb{R}^p. \tag{2}$$

Acima, a variável aleatória positiva U é denominada *fator de escala* com função de probabilidade h dependendo de um hiper-parâmetro $\boldsymbol{\tau}$ que foi associado ao peso nas caudas. Note que a distribuição normal p-variada é obtida quando h é o delta de Dirac centrado em 1.

Após anos de pesquisa, surgiram algumas formas de se combinar as novidades trazidas pelas duas distribuições anteriores, dentre as quais a recente proposta de [3]: as chamadas *distribuições misturas de escala assimétricas de normais* (SSMN na sigla em inglês). Segue adiante sua definição.

Definição 0.3. Um vetor aleatório $\mathbf{Y}$ p-dimensional segue uma *distribuição mistura de escala assimétrica da família normal* e escreve-se $\mathbf{Y} \sim \text{SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$ quando sua f.d.p. é

$$f(\mathbf{y}) = 2f_0(\mathbf{y})\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})),\ \mathbf{y} \in \mathbb{R}^p. \tag{3}$$

Na expressão acima f_0 é a f.d.p. dada em (2).

Essa nova classe combina os dois efeitos de assimetria e caudas pesadas, permitindo uma modelagem flexível, mas se defronta com a dificuldade computacional na estimação de seus parâmetros. A ideia inicial é a utilização da estimação por máxima verossimilhança (EMV), mas sua efetiva implementação torna-se mais simples por meio do algoritmo EM proposto em [2]. Para tanto, fez-se necessário enxergar as distribuições SSMN como uma classe de *modelos com dados incompletos* a partir das representações dadas adiante.

Proposição 0.1. Dado $\mathbf{Y} \sim \text{SSMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, temos que

$$\mathbf{Y} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\Big(U\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{-1}\left[\boldsymbol{\lambda}\left(\frac{U+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{U}\right)^{1/2}|T_0| + \Big(U\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\Big)^{1/2}\mathbf{T}_1\right]. \tag{4}$$

Acima, $T_0 \sim N_1(0,1), \mathbf{T}_1 \sim N_p(\mathbf{0}, \mathbf{I}_p)$ e $U \sim H(\,\cdot\,;\boldsymbol{\tau})$ são estatisticamente independentes. Tal representação estocástica gera a seguinte representação hierárquica para $\mathbf{Y}$:

$$\begin{aligned}
\mathbf{Y}|T=t, U=u &\sim N_p\left(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T)^{-1}\boldsymbol{\lambda}t, \boldsymbol{\Sigma}^{1/2}(u\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T)^{-1}\boldsymbol{\Sigma}^{1/2}\right) \\
T|U=u &\sim TN_{[0,+\infty)}\left(0, \frac{u+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{u}\right) \\
U &\sim H(\ .\ ;\boldsymbol{\tau})
\end{aligned} \tag{5}$$

Acima, $TN_{[0,+\infty)}$ é uma distribuição normal truncada no intervalo $[0,+\infty)$ e H é simplesmente a f.d.a. de U.

Mais especificamente, a estimação de todos os parâmetros desse modelo mostrou-se possível através das versões modificadas do algoritmo EM apresentadas em [7] (ECM) e [6] (ECME). Dada a complexidade envolvida na estimação do hiper-parâmetro $\boldsymbol{\tau}$, propôs-se em [3] o uso da segunda versão.

Observamos, no entanto, que para os três principais exemplos da classe de distribuições SSMN encontrados na literatura (T-Student normal assimétrica, Slash normal assimétrica e normal contaminada assimétrica) era factível adotar técnicas específicas capazes de acelerar o processo de estimação.

Esse resultado foi consequência da possibilidade de construir explicitamente em cada caso uma sequência de estimativas convergentes para a estimativa de máxima verossimilhança do hiper-parâmetro por meio do uso do algoritmo ECM e não mais do ECME, tornando o procedimento geral de estimação mais simples.

Na sequência apresentamos cada um dos exemplos mencionados, fazendo breve menção à técnica apropriada a cada um deles.

Definição 0.4. Dizemos que um vetor aleatório p-dimensional possui *distribuição T-Student normal assimétrica* e é escrito na forma $\mathbf{Y} \sim \mathrm{STN}_p\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big)$ quando sua f.d.p. é dada por

$$f\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big) = 2t_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\Big)\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right), \ \mathbf{y} \in \mathbb{R}^p. \tag{6}$$

Em (6), $t_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\Big) = \dfrac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\pi\nu)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left[1 + \dfrac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}}$ é a expressão da f.d.p. de uma distribuição T-Student p-variada com ν graus de liberdade.

Neste exemplo, usamos um método de aproximação presente em [8] para explicitar com uma margem de erro aceitável a sequência de estimativas que converge para a estimativa de máxima verossimilhança do parâmetro ν.

Definição 0.5. Dizemos que um vetor aleatório p-variado $\mathbf{Y}$ possui *distribuição Slash normal assimétrica* e escrevemos $\mathbf{Y} \sim \mathrm{SSN}_p\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big)$ quando sua f.d.p. é

$$f\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu\Big) = 2s_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right), \ \mathbf{y} \in \mathbb{R}^p. \tag{7}$$

Sendo $P\left(1; \nu + \frac{p}{2}, \frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$ a f.d.a. aplicada no ponto 1 de uma variável aleatória Gama $\left(\nu + \frac{p}{2}, \frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$, temos em (7) que a f.d.p. expressa por $s_p\Big(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{2^\nu\nu\Gamma\left(\nu+\frac{p}{2}\right)}{\pi^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]^{-\left(\nu+\frac{p}{2}\right)} P\left(1; \nu + \frac{p}{2}, \frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right)$ é da Slash simétrica.

Neste caso, estendemos uma mesma técnica de integração numérica utilizada em [5] para a estimação do parâmetro ν na Slash simétrica para o contexto assimétrico.

Definição 0.6. Denotamos um vetor aleatório p-dimensional com a *distribuição normal contaminada assimétrica* por $\mathbf{Y} \sim \mathrm{SCN}_p\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma\Big)$ quando sua f.d.p. é definida para todo $\mathbf{y} \in \mathbb{R}^p$ da forma expressa a seguir:

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu, \gamma) = 2\left[\nu\phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{\gamma}\right) + (1-\nu)\phi_p\left(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right]\Phi\left(\boldsymbol{\lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{y}-\boldsymbol{\mu})\right). \tag{8}$$

Neste exemplo, o hiper-parâmetro assume a forma $\boldsymbol{\tau} = (\nu, \gamma)$ e caracteriza uma mistura com fator de escala discreto.

A técnica empregada aqui para explicitar o hiper-parâmetro, também aplicada em [5] na distribuição normal contaminada simétrica – obtida fazendo $\boldsymbol{\lambda} = \mathbf{0}$ em (8), foi considerar tal distribuição como uma *mistura finita* (conforme [4]) de duas normais assimétricas.

Em todos os casos, mostraremos que as técnicas adotadas aceleram de duas a três vezes aproximadamente o processo de estimação e mantêm a consistência conforme ilustraremos em estudos de simulação.

Referências

[1] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[3] C. S. Ferreira, V. H. Lachos, and H. Bolfarine. Likelihood-based inference for multivariate skew scale mixtures of normal distributions. *AStA Advances in Statistical Analysis*, 100(4):421–441, 2016.

[4] S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[5] K. Lange and J. S. Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198, 1993.

[6] C. Liu and D. B. Rubin. The ecme algorithm: a simple extension of em and ecm with faster monotone convergence. *Biometrika*, pages 633–648, 1994.

[7] X. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[8] M. J. D. Powell. *Approximation theory and methods*. Cambridge university press, 1981.