Matlab Toolbox for Data Modeling Using GMM

Jan Švihlík,
Department of Computing and Control Engineering,
Institute of Chemical Technology, Prague,
e-mail: (jan.svihlik@vscht.cz).

Abstract

The presented toolbox contains several functions for data modeling using Gaussian Mixture Model (GMM) in its simplest form, i.e. sum of two Gaussian probability density functions (PDF). The parameters of GMM are estimated by using equation system derived by method of moments. GMM should model a signal and a noise in wavelet domain or in special cases also in spatial domain.

1 Gaussian mixture model

In this paper, image (signal) is modeled by the GMM in the wavelet domain. The GMM [1] is generally given by a mixture of a certain number of Gaussian PDFs with the variances σ_k and mean values μ_k

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x; \mu_k, \sigma_k^2), \tag{1}$$

where α_k are the proportions of the mixture. The parameters α_k satisfy the constraint $\sum_{k=1}^{K} \alpha_k = 1$. If $K = 2$, GMM is given by

$$p(x) = \alpha \mathcal{N}(x; \mu_1, \sigma_1^2) + (1 - \alpha)\mathcal{N}(x; \mu_2, \sigma_2^2). \tag{2}$$

The model given by (2) will be utilized in this paper for image modeling while mean value μ_1 and μ_2 are equal to zero.

The choice of two Gaussian PDF in GMM is a result of compromise between the solvability of the system of moment equations and the quality of the fit.

1.1 Derived system of equations

Let us consider image X in the wavelet domain. The second central theoretical moment [2] of X is given by

$$m_2(X) = \alpha_X \sigma_{1X}^2 + (1 - \alpha_X)\sigma_{2X}^2 \tag{3}$$

and the fourth moment runs as

$$m_4(X) = 3\alpha_X \sigma_{1X}^4 + 3(1 - \alpha_X)\sigma_{2X}^4, \tag{4}$$

where σ_{1X} denotes the first model variance corresponding to image and σ_{2X} represents the second model variance.

As a result, we derived two equations with three unknowns. Since we still have two equations with three unknowns, we exploit the the kurtosis κ_N [3] given by

$$\kappa_X = \frac{m_4(X)}{m_2^2(X)}. \tag{5}$$

From (3) and (4), we derive

$$\kappa_X = \frac{3\alpha_X \sigma_{1X}^4 + 3(1 - \alpha_X)\sigma_{2X}^4}{\alpha_X^2 \sigma_{1X}^4 + 2\alpha_X \sigma_{1X}^2 \sigma_{2X}^2 (1 - \alpha_X) + (1 - \alpha_X)^2 \sigma_{2X}^4}, \tag{6}$$

where the theoretical moments can be substituted by the sample moments $M_2(X)$ and $M_4(X)$ computed via $M_k(X) = \frac{1}{I}\sum_{i=1}^{I}(X_i - E(X))^k$. The first term after division should be equal to $\kappa_X \approx 3/\alpha_X$ ($for\ \alpha_X - 1 \to 0$). We empirically found that only this first term after division can be used for estimation of α_X. The variances σ_{1X} and σ_{2X} are estimated utilizing (3) and (4) in the following manner

$$\sigma_{1X}^2(3\alpha_X \sigma_{1X}^2 - 6\alpha_X m_2(X)) + \alpha_X m_4(X) - m_4(X) + 3m_2^2(X) = 0 \tag{7}$$

$$\sigma_{2X}^2 = \frac{m_2(X) - \alpha_X \sigma_{1X}^2}{1 - \alpha_X}. \tag{8}$$

The process of model parameters estimation may be simplified using the following equality $\sigma_{1X} = x_{0.999}/3$, where $x_{0.999}$ denotes the 99.9th percentile. The parameters estimation highly depends on the estimation quality of the sample moments.

2 Conclusion

The derived equation system is implemented in Matlab m-file *GMMEstPar.m*, whereas the toolbox contains also additional files such as *OptHistEval.m* (histogram optimization [4]), *JefDiv.m* (Jeffrey divergence evaluation [5]) and file for algorithm demonstration *demoGMM.m*. The presented algorithm is not so robust such as an expectation-maximization (EM) algorithm [6], but it is simple. Furthermore, it doesn't requires all data, but only measured two sample moments (second and fourth).

Acknowledgements

This work has been supported by the research project MSM 6046137306 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] A. Samé *et al.* (2007, Mar.), Mixture model-based signal denoising. *Advances in Data Anal. and Classification. 1(1)*, pp. 39-51.

[2] J. Švihlík (2009, Dec.), Modeling of Scientific Images Using GMM. *Radioengineering. 18(4)*, pp. 579-586.

[3] A. Pizurica, "Image denoising using wavelets and spatial context modeling," Ph.D. dissertation, Univ. Gent, Gent, Belgium, 2002.

[4] D. W. Scott (1979, Dec.), On optimal and data-based histograms. *Biometrika. 66(3)*, pp. 605-610

[5] P. Smith *et al.*, "Effective corner matching," in *Proc. of the Ninth BMVC 98,* Cambridge: Massachusetts Inst. of Technology, 1998.

[6] A. P. Dempster, N. M. Laird, D. B. Rubin, (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.* Series B (Methodological) *39(1)*, pp. 1-38