Why Mixture?

Tiancheng Li [1]

[1]Northwestern Polytechnical University

October 30, 2023

Abstract

From the most known Gaussian mixture to the cutting-edge multi-Bernoulli mixture of various forms, mixture offers a fundamental means to deal with uncertainties, which has led to a variety of appealing applications in the state estimation realm based on a single sensor or a sensor network. Like noise is often used to model unknown system input, one may use various hypotheses to deal with the uncertain state space model or data association. Meanwhile, consensus may be sought over the cross-correlated sensors. These all drive a need for representing the probability distribution by a mixture of properly weighted component distributions, which fuse the information gained from different models/hypotheses or from different sensors. This technical note presents information-theoretical results which answer how the averaging/mixture approach makes sense and how the fusing weights should be designed.

Why Mixture?

Tiancheng Li[a]

[a]*Key Laboratory of Information Fusion Technology, Northwestern Polytechnical University, Xi'an 710072, China.*
E-mail: t.c.li@nwpu.edu.cn; tiancheng.li1985@gmail.com

Abstract

From the most known Gaussian mixture to the cutting-edge multi-Bernoulli mixture of various forms, *mixture* offers a fundamental means to deal with uncertainties, which has led to a variety of appealing applications in the state estimation realm based on a single sensor or a sensor network. Like *noise* is often used to model unknown system input, one may use various hypotheses to deal with the uncertain state space model or data association. Meanwhile, consensus may be sought over the cross-correlated sensors. These all drive a need for representing the probability distribution by a mixture of properly weighted component distributions, which fuse the information gained from different models/hypotheses or from different sensors. This technical note presents information-theoretical results which answer how the averaging/mixture approach makes sense and how the fusing weights should be designed.

Keywords: Gaussian mixture; multi-Bernoulli mixture; consensus; arithmetic average; information fusion

---

Finite mixtures are flexible and powerful probabilistic modeling tools for both univariate and multivariate data, which have been well acknowledged and widely used for pattern recognition, machine learning, state estimation, etc. [1, 2] In the state estimation realm, the need for a mixture distribution may arise from stochastically switched models [3, 4, 5], multi-modal data/noise [6, 7, 8, 9] and data association uncertainty [10, 11, 12]. The most known mixture is the Gaussian mixture [13, 7], which consists of a finite number of Gaussian distributions. Recently, it has been further shown that the arithmetic average (AA) fusion which has provided a compelling approach to multi-target density fusion/consensus over sensor networks [14, 15, 16, 17, 18, 19, 20, 21] will also result in a mixture distribution.

In the Bayesian formulation, the state of interest $\mathbf{X}$ is considered random and the posterior is given in the manner of an estimate to the true probability distribution $p(\mathbf{X})$ of the state. The mixture distribution facilitates the closed-from Markov-Bayesian recursion greatly in two means: First, a mixture of conjugate priors is also conjugate and can approximate any kind of prior [22, 23]. Second, the linear fusion of a finite number of mixtures of the same parametric family remains a mixture of the same family. Therefore, the finite mixture has been one of the most important filter structures such as the known Gaussian mixture [13, 7], Student's-t mixture [24] and multi-Bernoulli mixture of various forms [25, 26, 27]. There are many other types of mixture models such as Watson mixture model [28] for axially symmetric data, inverted Beta mixture model [29] for non-symmetric data, von-Mises Fisher mixture model [30] for directional data (such as bearing measurements). Loosely speaking, the particle/Monte-Carlo method can also be viewed as a mixture of either variables (particle states) or distributions (Direct delta functions) [31]. These mixtures may convert to each other according to realistic needs or combine with each other resulting in hybrid mixtures such as Gaussian-Student's-t mixture [32] and Gaussian-uniform mixture [8].

In the mixture distribution, components/mixands are properly weighted and correspond to the information gained from different models/hypotheses or from different sensors. They joint approximate the true distribution $p(\mathbf{X})$ by their average/AA:

$$f_{\mathrm{AA}}(\mathbf{X}) = \sum_{s=1}^{S} w_s f_s(\mathbf{X}), \tag{1}$$

where $\mathbf{w} = [w_1, w_2, ..., w_S]^{\mathrm{T}}$ are nonnegative mixing/fusing weights which are typically normalized, namely $\mathbf{w}^{\mathrm{T}}\mathbf{1} = 1$, and $f_s(\mathbf{X}), s = 1, 2, \cdots, S$ are from the same parametric family in most cases.

Preprint submitted to techXiv. DOI: 10.36227/techrxiv.15082113 *July 30, 2021*

When the information is given in terms of variables (such as estimates of the number of targets), their average is a variable of smaller variance and so the benefit for averaging is obvious [33, 34]. However, for a mixture of general distributions (of the same family or not), it seems not so clear how good the averaging/mixture is and how the fusing weights should be designed. In this technical note, we study the exact divergence of the mixture distribution from the true/target distribution and briefly discuss principles for fusing/mixing weight design. These results are original according to the best of our knowledge and are expected to be useful for general mixture optimization and algorithm design.

1. Mixture Divergence

Theorem 1. *For a number of probability distributions $f_s(\mathbf{X}), s = 1, 2, \cdots, S$, the Kullback-Leibler (KL) divergence of any distribution $g(\mathbf{X})$ relative to their average $f_{AA}(\mathbf{X})$ is given as*

$$D_{KL}\left(f_{AA}\|g\right) = \sum_{s=1}^{S} w_s \big(D_{KL}(f_s\|g) - D_{KL}(f_s\|f_{AA})\big) \tag{2}$$

$$\leq \sum_{s=1}^{S} w_s D_{KL}(f_s\|g) \tag{3}$$

where the equation holds if and only if (iif) all fusing distributions $f_s, s = 1, 2, ..., S$ are identical.

Proof. The proof is straightforwardly derived as follows.

$$\begin{aligned}
D_{\mathrm{KL}}\left(f_{\mathrm{AA}}\|g)\right) =& D_{\mathrm{KL}}\left(\sum_{s=1}^{S} w_s f_s \middle\| g\right) \\
=& \int \sum_{s=1}^{S} w_s f_s(\mathbf{X}) \log \frac{f_{\mathrm{AA}}(\mathbf{X})}{g(\mathbf{X})} \delta\mathbf{X} \\
=& \sum_{s=1}^{S} w_s \left(\int f_s(\mathbf{X}) \log \frac{f_s(\mathbf{X})}{g(\mathbf{X})} \delta\mathbf{X} - \int f_s(\mathbf{X}) \log \frac{f_s(\mathbf{X})}{f_{\mathrm{AA}}(\mathbf{X})} \delta\mathbf{X} \right) \\
=& \sum_{s=1}^{S} w_s \left(D_{\mathrm{KL}}(f_s\|g) - D_{\mathrm{KL}}(f_s\|f_{\mathrm{AA}})\right) \\
\leq& \sum_{s=1}^{S} w_s D_{\mathrm{KL}}(f_s\|g)
\end{aligned}$$

where the equation holds iif $D_{\mathrm{KL}}(f_s\|f_{\mathrm{AA}}) = 0$, $s = 1, 2, ..., S$. □

Remark 1. When $g(\mathbf{X})$ is the true/target distribution $p(\mathbf{X})$, the above result indicates that *the average of the mixture fits the target distribution better than all component distributions on average.* This therefore provides an information-theoretic justification for distribution mixing/averaging, whether the information diversity is due to model/association uncertainty or sensing diversity. This is regardless of the mixing/fusion weights. optimized mixing weights will accentuate the benefit of fusion.

2. Mixing Weight

Mixing weights lie in the core of the mixture optimization and also play a key role in mixture reduction [35, 36, 14, 37, 38] according to case-specific needs. The naive weighting solution is the normalized uniform weights [39, 40], namely $\mathbf{w} = \mathbf{1}/S$. That is, all fusing distributions are treated equally which suits the case of fusing information from homogeneous sources. This is simple but does not distinguish the information of high quality from that of low.

More convincingly, the optimal solution should minimize $D_{\mathrm{KL}}\left(f_{\mathrm{AA}}\|p\right)$ in order to best fit the true/target distribution. That is, the optimal fusing weights $\mathbf{w}$ are determined as follows, c.f. (2)

$$\mathbf{w}_{\mathrm{opt}} = \arg\min_{\mathbf{w}} \sum_{s=1}^{S} w_s \big(D_{\mathrm{KL}}(f_s\|p) - D_{\mathrm{KL}}(f_s\|f_{\mathrm{AA}}) \big). \tag{4}$$

As shown in (4), the component that fits the target distribution better (corresponding to smaller $D_{\mathrm{KL}}(f_s\|p)$) and diverges more from the average (corresponding to greater $D_{\mathrm{KL}}(f_s\|f_{\mathrm{AA}})$) will be assigned with a greater fusing weight. However, the true/target distribution $p(\mathbf{X})$ is always unknown, so is $D_{\mathrm{KL}}(f_s\|p)$. It is also obvious that even if the true distribution $p(\mathbf{X})$ is available, the knowledge $D_{\mathrm{KL}}(f_s\|p) < D_{\mathrm{KL}}(f_i\|p)$, $\forall i \neq s$ does not necessarily result in $w_s = 1, w_i = 0, \forall i \neq s$. It also depends on $D_{\mathrm{KL}}(f_i\|f_s), \forall i \neq s$. That is, *the optimal fusion will not be fully dominated by the best component even if it is known.* In other words, c.f. Remark 1, we reach the following stronger claim on the *theoretically optimal solution*:

Remark 2. *When the fusing weights are properly designed, the average of the mixture may fit the target distribution better than the best component.*

An alternative solution is to resort to some functionally-similar divergences or metrics to assign higher weights to the components that fit the data better, namely having a higher likelihood. This *likelihood driven solution* is the key idea for weight updating in many mixture models/filters.

Another simplified alternative is ignoring the former part in (4) which will then be reduced approximately to the following suboptimal maximization problem

$$\mathbf{w}_{\mathrm{subopt}} = \arg\max_{\mathbf{w}} \sum_{s=1}^{S} w_s D_{\mathrm{KL}}(f_s\|f_{\mathrm{AA}}), \tag{5}$$

$$= \arg\max_{\mathbf{w}} \sum_{s=1}^{S} w_s \big(H(f_s, f_{\mathrm{AA}}) - H(f_s) \big), \tag{6}$$

where $H(f,g) := -\int f(\mathbf{X}) \log g(\mathbf{X}) \delta\mathbf{X}$ is the cross-entropy of distributions f and g, and $H(f) := H(f,f)$ is the differential entropy of distribution $f(\mathbf{X})$.

Remark 3. The suboptimal optimization given by (5)/(6) assigns a greater fusing weight to the distribution that diverges more from the others. This can be referred to as a *diversity preference solution.*

Nevertheless, one may design the fusing weights for some other purposes, e.g., in the context of seeking consensus over a peer-to-peer network [41, 42], they are typically designed for ensuring fast convergence [14, 15, 16, 17]. In the case of point estimate, a smaller variance is usually sought [43, 44, 14, 21].

3. Max-Min Optimization

Recall the divergence minimization that the AA fusion admits [45, 16]

$$f_{\mathrm{AA}} = \arg\min_{g} \sum_{s=1}^{S} w_s D_{\mathrm{KL}}(f_s\|g), \tag{7}$$

which holds for any nonnegative fusing weights $\mathbf{w}$ and is referred to as best fit of mixture (BFoM) [40].

Now, combining (7) with (5) yields joint optimization of the fusing form and fusing weights as follows

$$f_{\mathrm{AA}}(\mathbf{w}_{\mathrm{subopt}}) = \arg\max_{\mathbf{w}} \min_{g} \sum_{s=1}^{S} w_s D_{\mathrm{KL}}(f_s\|g). \tag{8}$$

This variational fusion problem (8) resembles that for geometric average (GA) fusion [46, 47], i.e.,

$$f_{\mathrm{GA}}(\mathbf{w}_{\mathrm{subopt}}) = \arg\max_{\mathbf{w}} \min_{g} \sum_{s=1}^{S} w_s D_{\mathrm{KL}}(g\|f_s), \tag{9}$$

where $f_{\mathrm{GA}}(\mathbf{w}) = C^{-1} \prod_{s=1}^{S} \big(f_s(\mathbf{X})\big)^{w_s}$ with $C = \int \prod_{s=1}^{S} \left(f_s(\mathbf{X})\right)^{w_s} \delta\mathbf{X}$.

It has actually been pointed out that the suboptimal fusion results for both variational fusion problems have equal KLD from/to the fusing distributions [46]. That is, $\forall i \neq j \in [1, S]$

$$D_{\mathrm{KL}}(f_i \| f_{\mathrm{AA}}(\mathbf{w}_{\mathrm{subopt}})) = D_{\mathrm{KL}}(f_j \| f_{\mathrm{AA}}(\mathbf{w}_{\mathrm{subopt}})), \tag{10}$$
$$D_{\mathrm{KL}}(f_{\mathrm{GA}}(\mathbf{w}_{\mathrm{subopt}}) \| f_i) = D_{\mathrm{KL}}(f_{\mathrm{GA}}(\mathbf{w}_{\mathrm{subopt}}) \| f_j), \tag{11}$$

which implies that the suboptimal fusion tends to revise all fusing estimators equivalently, resulting in a middle distribution where the AA and GA differs from each other in the direction of the KLD.

We reiterate that the above max-min solution (at least for the AA fusion) is suboptimal, which has ignored the minimization over $\sum_{s=1}^{S} w_s D_{\mathrm{KL}}(f_s \| p)$ and prefers diversity. Derivation for (11) has been earlier given in [43, 44] which is related to the Chernoff information [48] and has overlooked the potential benefit gained by the maximization in fitting the target distribution.

References

[1] G. McLachlan, D. Peel, Finite Mixture Models, Wiley Series in Probability and Statistics, Wiley, New York, USA, 2000.

[2] J.-M. Marin, K. Mengersen, C. P. Robert, Bayesian modelling and inference on mixtures of distributions, in: D. Dey, C. Rao (Eds.), Bayesian Thinking, Vol. 25 of Handbook of Statistics, Elsevier, 2005, pp. 459–507. doi:https://doi.org/10.1016/S0169-7161(05)25016-2.

[3] X. R. Li, V. P. Jilkov, Survey of maneuvering target tracking. part v. multiple-model methods, IEEE Trans. Aerosp. Electron. Syst. 41 (4) (2005) 1255–1321. doi:10.1109/TAES.2005.1561886.

[4] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models, Springer Series in Statistics, Springer New York, New York, NY, 2006.

[5] T. M. Fragoso, W. Bertoli, F. Louzada, Bayesian model averaging: A systematic review and conceptual classification, Int Stat Rev 86 (1) (2018) 1–28. doi:10.1111/insr.12243.

[6] R. L. Streit, T. E. Luginbuhl, Maximum likelihood method for probabilistic multihypothesis tracking, in: O. E. Drummond (Ed.), Signal and Data Processing of Small Targets 1994, Vol. 2235, International Society for Optics and Photonics, SPIE, 1994, pp. 394 – 405. doi:10.1117/12.179066.

[7] L. Pishdad, F. Labeau, Analytic MMSE bounds in linear dynamic systems with Gaussian mixture noise statistics, arXiv:1506.07603.

[8] P. Coretto, C. Hennig, Maximum likelihood estimation of heterogeneous mixtures of Gaussian and uniform distributions, Journal of Statistical Planning and Inference 141 (1) (2011) 462–473. doi:https://doi.org/10.1016/j.jspi.2010.06.024.

[9] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Information Fusion 59 (2020) 103–126. doi:10.1016/j.inffus.2020.01.011.

[10] D. Reid, An algorithm for tracking multiple targets, IEEE Transactions on Automatic Control 24 (6) (1979) 843–854. doi:10.1109/TAC.1979.1102177.

[11] R. P. S. Mahler, Advances in Statistical Multisource-Multitarget Information Fusion, Artech House, Norwood, MA, USA, 2014.

[12] B.-N. Vo, M. Mallick, Y. Bar-shalom, S. Coraluppi, R. Osborne, R. Mahler, B.-T. Vo, Multitarget tracking, in: Wiley Encyclopedia of Electrical and Electronics Engineering, John Wiley & Sons, 2015. doi:10.1002/047134608X.W8275.

[13] H. Sorenson, D. Alspach, Recursive Bayesian estimation using Gaussian sums, Automatica 7 (4) (1971) 465–479. doi:10.1016/0005-1098(71)90097-5.

[14] T. Li, J. Corchado, S. Sun, Partial consensus and conservative fusion of Gaussian mixtures for distributed PHD fusion, IEEE Trans. Aerosp. Electron. Syst. 55 (5) (2019) 2150–2163.

[15] T. Li, F. Hlawatsch, A distributed particle-PHD filter using arithmetic-average fusion of Gaussian mixture parameters, Information Fusion 73 (2021) 111–124. doi:https://doi.org/10.1016/j.inffus.2021.02.020.

[16] T. Li, Z. Liu, Q. Pan, Distributed Bernoulli filtering for target detection and tracking based on arithmetic average fusion, IEEE Signal Processing Letters 26 (12) (2019) 1812–1816. doi:10.1109/LSP.2019.2950588.

[17] T. Li, X. Wang, Y. Liang, Q. Pan, On arithmetic average fusion and its application for distributed multi-Bernoulli multitarget tracking, IEEE Trans. Signal Process. 68 (2020) 2883–2896.

[18] W. Yi, G. Li, G. Battistelli, Distributed multi-sensor fusion of PHD filters with different sensor fields of view, IEEE Trans. Signal Process. 68 (2020) 5204–5218. doi:10.1109/TSP.2020.3021834.

[19] L. Gao, G. Battistelli, L. Chisci, Multiobject fusion with minimum information loss, IEEE Signal Process. Lett. 27 (2020) 201 – 205.

[20] K. Da, T. Li, Y. Zhu, Q. Fu, Gaussian mixture particle jump-Markov-CPHD fusion for multitarget tracking using sensors with limited views, IEEE Trans. Signal Inform. Process. Netw. 6 (2020) 605–616.

[21] K. Da, T. Li, Y. Zhu, H. Fan, Q. Fu, Recent advances in multisensor multitarget tracking using random finite set, Front Inform Technol Electron Eng. 22 (1) (2021) 5–24.

[22] S. R. Dalal, W. J. Hall, Approximating priors by mixtures of natural conjugate priors, Journal of the Royal Statistical Society: Series B (Methodological) 45 (2) (1983) 278–286. doi:10.1111/j.2517-6161.1983.tb01251.x.

[23] P. Diaconis, Y. D., Quantifying prior opinion, Report Number: EFS NSF207.

[24] D. Peel, G. McLachlan, Robust mixture modelling using the t distribution, Stat. Comput. 10 (4) (2000) 339–348.
[25] B.-T. Vo, B.-N. Vo, Labeled random finite sets and multi-object conjugate priors, IEEE Trans. Signal Process. 61 (13) (2013) 3460–3475. doi:10.1109/TSP.2013.2259822.
[26] J. L. Williams, Marginal multi-Bernoulli filters: RFS derivation of MHT, JIPDA, and association-based MeMBer, IEEE Trans. Aerosp. Electron. Syst. 51 (3) (2015) 1664–1687. doi:10.1109/TAES.2015.130550.
[27] A. F. García-Fernández, J. L. Williams, K. Granström, L. Svensson, Poisson multi-Bernoulli mixture filter: Direct derivation and implementation, IEEE Trans. Aerosp. Electron. Syst. 54 (4) (2018) 1883–1901.
[28] J. Taghia, A. Leijon, Variational inference for watson mixture model, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (9) (2016) 1886–1900. doi:10.1109/TPAMI.2015.2498935.
[29] Y. Lai, X. Ma, Y. Xu, Y. Ling, C. Du, J. Du, Y. Zhang, Y. Ping, Positive data modeling using a mixture of mixtures of inverted beta distributions, IEEE Access 7 (2019) 38146–38156. doi:10.1109/ACCESS.2019.2906669.
[30] J. Taghia, Z. Ma, A. Leijon, Bayesian estimation of the von-mises fisher mixture model with variational inference, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (9) (2014) 1701–1715. doi:10.1109/TPAMI.2014.2306426.
[31] P. Moral, Mean Field Simulation for Monte Carlo Integration, Chapman and Hall/CRC, 2013.
[32] Y. Huang, Y. Zhang, Y. Zhao, J. A. Chambers, A novel robust Gaussian-student's t mixture distribution based Kalman filter, IEEE Transactions on Signal Processing 67 (13) (2019) 3606–3620. doi:10.1109/TSP.2019.2916755.
[33] T. Li, F. Hlawatsch, P. M. Djuríc, Cardinality-consensus-based PHD filtering for distributed multitarget tracking, IEEE Signal Process. Lett. 26 (1) (2019) 49–53.
[34] T. Li, H. Fan, J. García, J. M. Corchado, Second-order statistics analysis and comparison between arithmetic and geometric average fusion: Application to multi-sensor target tracking, Information Fusion 51 (2019) 233 – 243.
[35] D. J. Salmond, Mixture reduction algorithms for target tracking in clutter, in: Proc. SPIE, Vol. 1305, 1990, pp. 434–445. doi:10.1117/12.21610.
[36] D. F. Crouse, P. Willett, K. Pattipati, L. Svensson, A look at Gaussian mixture reduction algorithms, in: 14th Int. Conf. Inf. Fusion, 2011, pp. 1–8.
[37] H. Kim, K. Granström, L. Gao, G. Battistelli, S. Kim, H. Wymeersch, 5g mmwave cooperative positioning and mapping using multi-model PHD filter and map fusion, IEEE Trans. Wireless Commun. 19 (6) (2020) 3782–3795.
[38] R. K. Ramachandran, N. Fronda, G. Sukhatme, Resilience in multi-robot multi-target tracking with unknown number of targets through reconfiguration, IEEE Transactions on Control of Network SystemsIn press. doi:10.1109/TCNS.2021.3059794.
[39] T. Li, V. Elvira, H. Fan, J. M. Corchado, Local-diffusion-based distributed SMC-PHD filtering using sensors with limited sensing range, IEEE Sensors Journal 19 (4) (2019) 1580–1589.
[40] T. Li, K. Da, Z. Liu, X. Wang, Y. Liang, Best fit of mixture for computationally efficient poisson multi-bernoulli mixture filtering, DOI: 10.36227/techrxiv.12351710.Under revision in IEEE TSP.
[41] L. Xiao, S. Boyd, Fast linear iterations for distributed averaging, Syst. Control. Lett. 53 (1) (2004) 65–78. doi:http://doi.org/10.1016/j.sysconle.2004.02.022.
[42] R. Olfati-Saber, J. A. Fax, R. M. Murray, Consensus and cooperation in networked multi-agent systems, Proc. IEEE 95 (1) (2007) 215–233. doi:10.1109/JPROC.2006.887293.
[43] M. B. Hurley, An information theoretic justification for covariance intersection and its generalization, in: Proc. FUSION 2002, Annapolis, MD, USA, 2002, pp. 505–511. doi:10.1109/ICIF.2002.1021196.
[44] S. J. Julier, An empirical study into the use of chernoff information for robust, distributed fusion of Gaussian mixture models, in: Proc. FUSION 2006, Florence, Italy, 2006. doi:10.1109/ICIF.2006.301755.
[45] K. Da, T. Li, Y. Zhu, H. Fan, Q. Fu, Kullback-Leibler averaging for multitarget density fusion, in: Proc. DCAI 2019, Avila, Spain, 2019, pp. 253–261.
[46] F. Nielsen, An information-geometric characterization of Chernoff information, IEEE Signal Processing Letters 20 (3) (2013) 269–272. doi:10.1109/LSP.2013.2243726.
[47] M. Üney, J. Houssineau, E. Delande, S. J. Julier, D. E. Clark, Fusion of finite set distributions: Pointwise consistency and global cardinality 55 (6) (2019) 2759–2773, published online. doi:10.1109/TAES.2019.2893083.
[48] T. M. Cover, J. A. Thomas, Elements of Information Theory, John Wiley & Sons, Ltd, 2001. doi:10.1002/0471200611.