Edgeworth-expanded Topographic Map Formation

Marc M. Van Hulle
K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg, Herestraat 49, B-3000 Leuven, BELGIUM
E-mail: marc@neuro.kuleuven.ac.be

Abstract - *We introduce a new learning algorithm for topographic map formation with Edgeworth-expanded Gaussian activation kernels. First, we consider mixtures of Edgeworth-expanded Gaussians for modeling the input density, and derive a simple closed form solution for estimating the kernel parameters based on weighted moment matching. Then, the topographic map formation algorithm is introduced which is based on the batch map algorithm and weighted moment matching.*

Key words - Gaussian mixture modeling, Edgeworth expansion, constrained log-likelihood maximization, batch map

1 Introduction

Several unsupervised learning algorithms have been devised that develop topographically-organized maps of Gaussian mixture densities (for references, see [10]). Unifying accounts of the homoscedastic (equal-variance) case have been introduced by Graepel and co-workers [5], and Heskes [6]. The former authors adopted a statistical physics approach, showing the connection between different classes of Gaussian kernel-based topographic map formation algorithms and, as a limiting case, the batch map version of the Self-Organizing Map (SOM) algorithm [8]. Heskes showed the connection between minimum distortion topographic map formation and maximum likelihood homoscedastic Gaussian mixture density modeling. Another approach is to minimize the Kullback-Leibler divergence, an idea that has been introduced in kernel-based topographic map formation by Benaim and Tomasini [3], using homoscedastic Gaussians, and extended more recently by Yin and Allinson [12] to heteroscedastic (different-variance) Gaussians. A unifying account of the heteroscedastic case has been introduced by Van Hulle [11], showing the connection between minimum distortion topographic map formation, maximum likelihood heteroscedastic Gaussian mixture density modeling and Kullback-Leibler divergence.
In principle, kernels other than Gaussians could be used in topographic map formation (*e.g.*, see [10]). In this article, we suggest the Edgeworth-expanded Gaussian kernel, which consists of a Gaussian kernel multiplied by a series of Hermite polynomials of increasing order, and of which the coefficients can be estimated through higher-than-second-order moment matching. The Edgeworth expansion of the univariate Gaussian density became popular in the neural network community when it was introduced in Independent Component Analysis (ICA) [1, 4] and projection pursuit [7].
The article is organized as follows. We first recall the univariate Edgeworth expansion and

then, as a first novelty, derive its extension to univariate Edgeworth-expanded Gaussian mixture density modeling, based on constrained log-likelihood maximization. Next, we consider the multivariate Edgeworth-expansion and multivariate Edgeworth-expanded Gaussian mixture density modeling. Finally, as a second novelty, we introduce a topographic map formation algorithm for Edgeworth-expanded Gaussian kernels.

2 Univariate Edgeworth Expansion

The Edgeworth series expansion of the scalar density $p(v)$, $v \in V \subseteq \Re$, around its best Gaussian estimate ϕ_p (*i.e.*, with the same mean μ and standard deviation σ as p) is [2]:

$$p(v) \approx \phi_p(v)\left(1+\frac{1}{3!}\kappa_3 H_3(v)+\frac{1}{4!}\kappa_4 H_4(v)+\frac{10}{6!}\kappa_3^2 H_6(v)+\frac{1}{5!}\kappa_5 H_5(v)+\dots\right), \tag{1}$$

with κ_i the ith *standardized* cumulant, and H_i the Hermite polynomial of order i, $H_3(v) = z^3-3z, H_4 = z^4-6z^2+3, H_5 = z^5-10z^3+15z, H_6 = z^6-15z^4+45z^2-15$, with z the standardized scalar $z=\frac{v-\mu}{\sigma}$. It can be verified that the Edgeworth expansion cannot be made arbitrarily good by including terms of higher and higher orders. Usually, the third- or fourth-order (Hermite polynomial) expansion suffices.

3 Univariate Edgeworth-expanded Gaussian Mixture

We now consider the homogeneous mixture density estimate $p(v) \approx \frac{1}{N}\sum_i p(v|i, w_i, \sigma_i, \kappa_{3i}, \kappa_{4i})$, in which we take for $p(v|i, w_i, \sigma_i, \kappa_{3i}, \kappa_{4i})$ the Edgeworth expanded kernels eq. (1), up to the fourth-order. In order to determine the cumulants, consider first the following constrained optimization problem with which the moments can be determined:

$$\min\{F\} \text{ subject to}: \ (\kappa_{3i})^2-(\frac{\mu_{3i}}{\sigma_i^3})^2 \leq 0 \text{ and } (\kappa_{4i})^2-(\frac{\mu_{4i}-3\sigma_i^4}{\sigma_i^4})^2 \leq 0, \quad \forall i, \tag{2}$$

where we have added the subscript i to denote kernel i, and where the constraints correspond to the 3rd and 4th standardized cumulants of kernel i, just as if kernel i would be the only kernel. We solve the optimization problem with the Lagrange multipliers technique:

$$\begin{aligned} F_{\text{constr}} &= -\frac{1}{M}\sum_n \log\frac{1}{N}\sum_i p(v|i, w_i, \sigma_i, \kappa_{3i}, \kappa_{4i}) + \sum_i \lambda_{3i}\left((\kappa_{3i})^2-(\frac{\mu_{3i}}{\sigma_i^3})^2\right) \\ &\quad + \sum_i \lambda_{4i}\left((\kappa_{4i})^2-(\frac{\mu_{4i}-3\sigma_i^4}{\sigma_i^4})^2\right) \end{aligned} \tag{3}$$

with the λ_{3i} and $\lambda_{4i}, \forall i$, the Lagrange multipliers. The result leads to the following fixed point update rules, used in an Expectation-Maximization (EM) format:

$$\begin{aligned} \mu_{1i} &= \frac{\sum_n P(i|v^n)v^n}{\sum_n P(i|v^n)}, & \mu_{2i} &= \frac{\sum_n P(i|v^n)(v^n-w_i)^2}{\sum_n P(i|v^n)}, \\ \mu_{3i} &= \frac{\sum_n P(i|v^n)(v^n-w_i)^3}{\sum_n P(i|v^n)}, & \mu_{4i} &= \frac{\sum_n P(i|v^n)(v^n-w_i)^4}{\sum_n P(i|v^n)}, \end{aligned}$$

with $\mu_{1i} \equiv w_i$, $\mu_{2i} \equiv \sigma_i^2$ (we have changed the notation of our kernel parameters so as to put in the Edgeworth expansion framework). Note that we get the same result when reversing

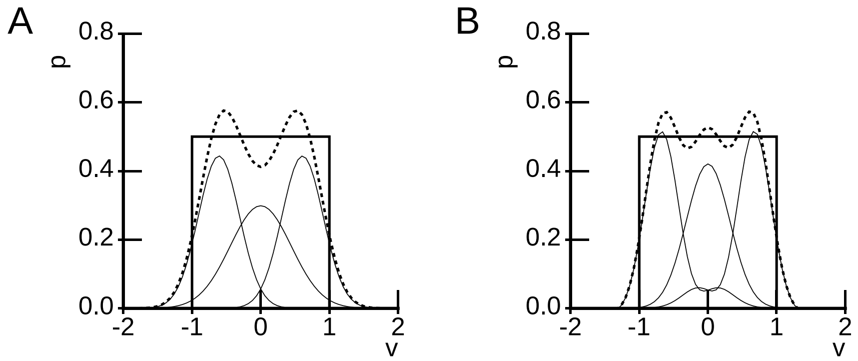

Figure 1: (A) Gaussian mixture density estimate (dashed line) (A) and fourth-order Edgeworth-expanded mixture density estimate (dashed line) (B) for the case of 3 kernels (thin full lines) and the uniform input distribution $[-1, 1]$ (thick full line).

the inequality signs in eq. (2). The cumulants are thus determined in a similar way as the first two cumulants in the Gaussian mixture case (*i.e.*, weighted moment matching).

As an example, we consider the uniform distribution $[-1, 1]$ modeled using $N = 3$ Gaussian kernels and 3 fourth-order Edgeworth-expanded Gaussian kernels (Fig. 1). We observe the smaller ripple in the Edgeworth-expanded case. The mean squared error (MSE) between the correct distribution and the estimate is 4.59×10^{-3}, for the Gaussian and 3.20×10^{-3} for the Edgeworth-expanded case.

4 Multivariate Edgeworth Expansion

Let $\mathbf{v} = [v_1, \ldots, v_d] \in V \subseteq \mathcal{R}^d$ be a random vector drawn from the density $p(\mathbf{v})$. The Edgeworth expansion of $p(\mathbf{v})$, up to order five about its best normal estimate, is given by [2]:

$$p(\mathbf{v}) \approx \phi_p(\mathbf{v}) \left(1 + \sum_{i,j,k} \frac{1}{3!} \kappa_{i,j,k} H_{ijk}(\mathbf{v}) + \sum_{i,j,k,l} \frac{1}{4!} \kappa_{i,j,k,l} H_{ijkl}(\mathbf{v}) + \sum_{i,j,k,l,p,q} \frac{1}{72!} \kappa_{i,j,k} \kappa_{l,p,q} H_{ijklpq}(\mathbf{v}) \ldots \right), \tag{4}$$

with H_{ijk} the ijk-th Hermite polynomial, with i, j, k the corresponding input dimensions, $i, j, k \in \{1, \ldots, d\}$, and $\kappa_{i,j,k}$ the corresponding standardized cumulant, $\kappa_{i,j,k} = \frac{\kappa_{ijk}}{\sqrt{\sigma_i^2 \sigma_j^2 \sigma_k^2}}$, with κ_{ijk} the third cumulant over input dimensions i, j, k, and where the sum over all combinations i, j, k is considered, and H_{ijkl} the $ijkl$-th Hermite polynomial over input dimensions i, j, k, l, and $\kappa_{i,j,k,l}$ the corresponding standardized cumulant, $\kappa_{i,j,k,l} = \frac{\kappa_{ijkl} - \sigma_{ij}\sigma_{kl}[3]}{\sqrt{\sigma_i^2 \sigma_j^2 \sigma_k^2 \sigma_l^2}}$, with κ_{ijkl} the fourth cumulant over input dimensions i, j, k, l, $\sigma_{ij}\sigma_{kl}[3]$ the sum over the 3 partitions of 4 indices and σ_{ij} the covariance between v_i and v_j. Finally, the cumulants are obtained from the moments using the formula of McCullagh (1987).

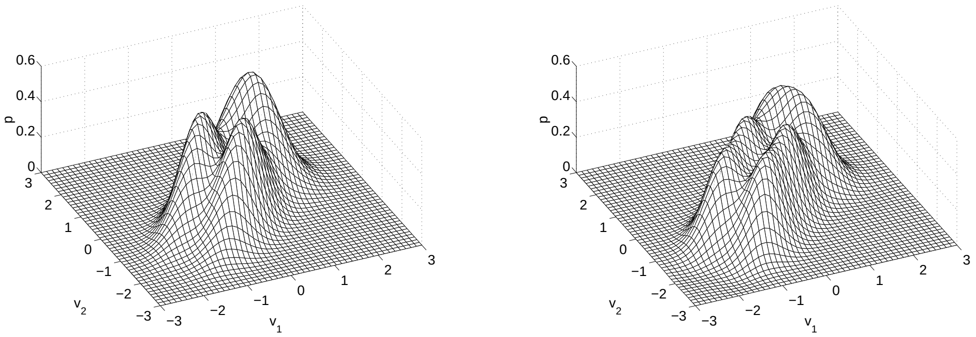

Figure 2: *Left panel:* Mixture density estimate of a 2×1 rectangular uniform density, rotated over 45 deg, using 3 heteroscedastic Gaussian kernels. *Right panel:* Idem but using 3 Edgeworth-expanded Gaussian kernels.

5 Multivariate Edgeworth-expanded Gaussian Mixture

The weighted moments are estimated in the same way as in the univariate case, thus, also as the solutions of a constrained log-likelihood procedure. As an example, consider a 2×1 rectangular uniform density rotated over 45 deg with 100k data points drawn from it. We take 3 kernels of which the centers are initialized randomly by sampling the input density. We further initialize with a diagonal covariance matrix with diagonal elements equal to 0.5; all initial third and fourth moments are zero. The result is shown in Fig. 2. We observe that 2 Edgeworth-expanded kernels are stretched along the long axis with almost flat peaks, which indicates a substantial fourth-order moment contribution along this axis. In addition to an improved capturing of the overall shape of the distribution, the MSEs are 0.0110 and 0.00936 for the Gaussian and Edgeworth-expanded cases, respectively.

6 Topographic Map Formation

One possibility for topographic map formation which comes to mind is to start with Heskes' update rule for the Gaussian kernel centers [6]:

$$\mathbf{w}_i = \frac{\sum_n \sum_j p(j|\mathbf{v}^n)\lambda(i,j)\mathbf{v}^n}{\sum_n \sum_j p(j|\mathbf{v}^n)\lambda(i,j)}, \tag{5}$$

with $\mathbf{w}_i = [w_{1i}, \ldots, w_{di}]$, and $\lambda(i,j)$ the usual neighborhood function. It is tempting to update the kernel's second moments in a similar way, inspired by the second moments in Gaussian mixture density modeling, *e.g.*, for the variance along the first dimension:

$$\sigma_{1i}^2 = \frac{\sum_n \sum_j p(j|\mathbf{v}^n)\lambda(i,j)(v_1^n - w_{1i})^2}{\sum_n \sum_j p(j|\mathbf{v}^n)\lambda(i,j)}. \tag{6}$$

However, as shown in [11], when deriving the update rule in a maximum log-likelihood format, the presence of the neighborhood makes that we no longer have a closed form solution for the second moments. Only when the neighborhood vanishes this is the case (which actually reverts to regular Gaussian mixture density modeling). Furthermore, even when we would

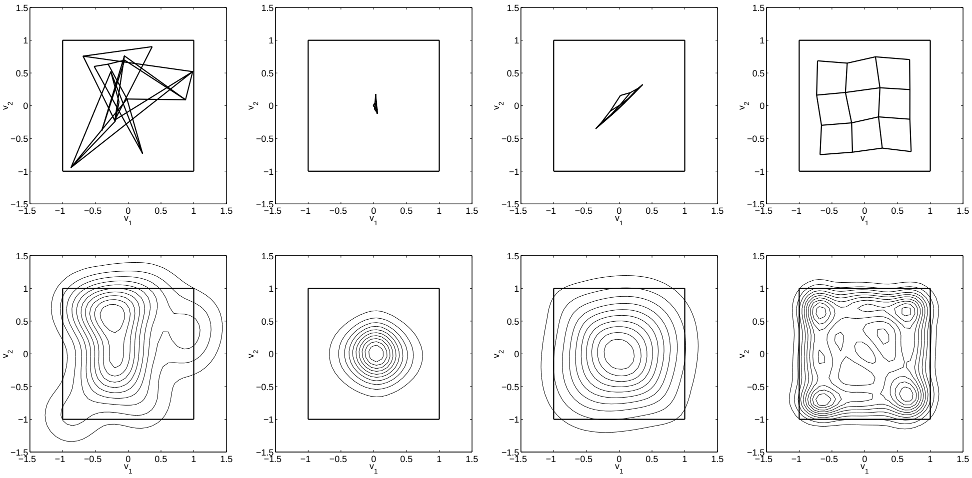

Figure 3: *Top row:* Topographic map formation of 4×4 lattice given 10k data points taken from a uniform square distribution $[-1,1]^2$ (square boxes). Snapshots for epochs 0,1,10,100 (from left to right). *Bottom row:* Corresponding density estimates shown as contour plots (10 equidistant levels).

consider this possibility as a heuristic, the Gaussian kernels' ranges will grow large during the contraction phase of the topographic map formation process, so as to span as much as possible the input distribution, but the algorithm cannot recover from these large spans and lumped kernel centers, even when the neighborhood has completely vanished. Clearly this seems to be a local, non-optimal log-likelihood solution. In order to remedy this, we consider a heuristic procedure in which we update the kernel centers according to Kohonen's batch map algorithm [8]:

$$\mathbf{w}_i = \frac{\sum_n \lambda(i^*(\mathbf{v}^n), i)\mathbf{v}^n}{\sum_n \lambda(i^*(\mathbf{v}^n), i)}, \quad \forall i, \tag{7}$$

but in which we select the winner according to the Edgeworth-expanded kernel that is most active: $i^*(\mathbf{v}^n) = \arg\max_i\{p(\mathbf{v}^n|i)\}$. We then update the other parameters of the kernels as done in the Edgeworth-expanded Gaussian mixture case. In this way, we are able to both develop a topographic map and to perform mixture density modeling, but constrained by the locations of the kernel centers.

As an example, consider the uniform square distribution $[-1,1]^2$, from which 10k data points are taken, and a $N = 4 \times 4$ lattice. The kernel centers are initialized by sampling the input distribution, and the covariance matrix is initialized by adopting a diagonal one with non-zero elements equal to 0.1^2; all initial third and fourth moments are equal to zero. A Gaussian neighborhood function $\lambda(.)$ is chosen with initial range $\sigma_\lambda(0) = 2$, and which is decreased exponentially over $t_{\max} = 100$ epochs: $\sigma_\lambda(t) = 2\exp(-2t/t_{\max})$. At each epoch t, we perform one step of the batch map algorithm, and then iterate between the expectation- and maximization steps in the EM stage to determine the second, third and fourth moments until convergence (usually a few runs). A few snap shots of the map formation process are shown in Fig. 3 together with the density estimates.

Finally, since the computational complexity rapidly increases with dimensionality, we could, as a simplification, consider only the third and fourth-order expansions along the dimensions with the largest second moment(s) of each kernel.

7 Conclusion

We have introduced two novelties in this article. First, we derived a simple closed form solution for estimating the parameters of mixtures of Edgeworth-expanded Gaussian kernels based on weighted moment matching. Second, we introduced a new topographic map formation algorithm for Edgeworth-expanded Gaussian activation kernels, based on the batch map algorithm and weighted moment matching.

Acknowledgments

The author is supported by the Belgian Fund for Scientific Research – Flanders (G.0248.03, G.0234.04), the Flemish Regional Ministry of Education (Belgium) (GOA 2000/11), the Belgian Science Policy (IUAP P5/04), and the European Commission (NEST-2003-012963).

References

[1] S.-I. Amari, A. Cichocki, and H. Yang (1996), A new learning algorithm for blind signal separation, In *Advances in Neural Processing Systems* (Vol. 8), D.S. Touretzky, M. Mozer and M. Hasselmo (Eds.), Cambridge, MA: MIT Press, pp. 757-763.

[2] O.E. Barndorff-Nielsen, and D.R. Cox (1989), *Inference and Asymptotics*, London, Chapman and Hall.

[3] M. Benaim and L. Tomasini (1991), Competitive and self-organizing algorithms based on the minimization of an information criterion, *Proc. ICANN'91*, pp. 391-396.

[4] P. Comon (1994), Independent component analysis – a new concept?, *Signal Processing*, **vol. 36**(3), pp. 287-314.

[5] T. Graepel, M. Burger, and K. Obermayer (1998), Self-organizing maps: Generalizations and new optimization techniques, *Neurocomputing*, **vol. 21**, pp. 173-190.

[6] T. Heskes (2001), Self-organizing maps, vector quantization, and mixture modeling, *IEEE Trans. Neural Networks*, **vol. 12**(6), pp. 1299-1305.

[7] M. Jones, and R. Sibson (1987), What is projection pursuit?, *J. of the Royal Statistical Society A*, **vol. 150**, pp. 1-36.

[8] T. Kohonen (1995), *Self-organizing maps,* Heidelberg: Springer.

[9] P. McCullagh (1987), *Tensor methods in statistics.* London: Chapman and Hall.

[10] M.M. Van Hulle (2002), Joint entropy maximization in kernel-based topographic maps, *Neural Computat.*, **vol. 14**(8), pp. 1887-1906.

[11] M.M. Van Hulle (2005), Maximum likelihood topographic map formation, *Neural Computat.*, **vol. 17**(3), pp. 503-513.

[12] H. Yin and N.M. Allinson (2001), Self-organizing mixture networks for probability density estimation, *IEEE Trans. Neural Networks*, **vol. 12**, pp. 405-411.