IDIAP RESEARCH REPORT

Conditional Gaussian Mixtures

Todd A. Stephenson [a,b]

IDIAP–RR 03-11

January 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

a Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
b Swiss Federal Institute of Technology Lausanne (EPFL)

IDIAP Research Report 03-11

Conditional Gaussian Mixtures

Todd A. Stephenson

January 2003

Abstract. I show how conditional Gaussians, whose means are conditioned by a random variable, can be estimated and their likelihoods computed. This is based upon how regular Gaussians have their own parameters and likelihood computed. After explaining how to estimate the parameters of Gaussians and conditional Gaussians, I explain how to calculate their likelihoods even if there are missing elements in the data or, in the case of the conditional Gaussian, even if the conditioning variable is missing.

Acknowledgements: Todd A. Stephenson is supported by the Swiss National Science Foundation under the grant BN_ASR (20-64172.00).

1 Parameter Estimation

1.1 Gaussians

In machine learning, the Gaussian is a common distribution for modeling a wide variety of data. It consists of two parameters: the mean vector μ and the covariance matrix Σ, which are referred to as the first moment and the second moment, respectively. If data X is distributed in such a manner, we note it as:

$$X \sim \mathcal{N}(\mu_X, \Sigma_X) \tag{1}$$

If μ_X and Σ_X are not known but if we do have data $X = \{x_1, \ldots, x_N\}$ drawn from this distribution, we can estimate their values:

$$\mu_X = E(X) \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

$$\Sigma_X = E((\mu_X - E(X))^2) \approx S_X = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T, \tag{3}$$

where (3) is a biased estimator of Σ_X. Alternatively, we can have an unbiased estimator of Σ_X:

$$S_X = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T, \tag{4}$$

where (4) divides by $(N-1)$ instead of by N. The same result for S_X in (3) can be obtained by:

$$\begin{aligned} S_X &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T - \frac{1}{N}\sum_{i=1}^{N} 2x_i\bar{x}^T + \frac{1}{N}\sum_{i=1}^{N}\bar{x}\bar{x}^T \\ &= \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T - \bar{x}\bar{x}^T. \end{aligned} \tag{5}$$

The advantage of (5) over (3) is that both μ_X and Σ_X can be estimated in the same pass over the data, instead of two.

1.2 Conditional Gaussians

The conditional Gaussian is a less common distribution used for modeling the dependency between data X and data Y. If we define X as being dependent upon Y, then we have a distribution whose first moment is no longer the mean of X but, rather, the mean of X conditioned by the given value y from Y: $u_X = \mu_X + B_X\, y$, using regressions in B_X upon y. Its second moment is still the variance Σ_X (albeit, estimated differently, as explained below):

$$X \sim \mathcal{N}(u_X, \Sigma_X). \tag{6}$$

Hence, whereas a Gaussian has two parameters to estimate, μ and Σ, a conditional Gaussian has three parameters: μ, B, and Σ. These parameters are learned in a two-stage procedure. First, we need to estimate the joint mean and joint covariance of X and Y. Let W be the combination of data X and

Y (that is, for each pair of data vectors x and y, w is the data vector which is the concatenation of the two). Then we estimate μ_W and Σ_W:

$$\mu_W \approx \bar{w} = \frac{1}{N}\sum_{i=1}^{N} w \tag{7}$$

$$\Sigma_W \approx S_W = \frac{1}{N}\sum_{i=1}^{N}(w-\bar{w})(w-\bar{w})^T = \frac{1}{N}\sum_{i=1}^{N} ww^T - \bar{w}\bar{w}^T \tag{8}$$

Let us partition $\bar{w}$ according to the portion obtained from X and that obtained from Y:

$$\bar{w} = \begin{bmatrix} \bar{w}\{X\} \\ \bar{w}\{Y\} \end{bmatrix} \tag{9}$$

Let us also partition S_W, according to that related only to X, that related only to Y and those related jointly to X and Y, as follows

$$S_W = \begin{bmatrix} S_W\{X^2\} & S_W\{XY\} \\ S_W\{YX\} & S_W\{Y^2\} \end{bmatrix} \tag{10}$$

and let $[S_W\{Y^2\}]^-$ be the pseudo-inverse of $S_W\{Y^2\}$. Then the parameters μ_X, B_X, and Σ_X for the conditional Gaussian are estimated as:

$$B_X \approx \hat{B}_X = S_W\{XY\}[S_W\{Y^2\}]^- \tag{11}$$

$$\mu_X \approx \bar{x}_X = \bar{w}\{X\} - \hat{B}_X\bar{w}\{Y\} \tag{12}$$

$$\Sigma_X \approx S_X = S_X - \hat{B}_X S_W\{YX\}. \tag{13}$$

Note that while only the first moment, u_X, is conditioned upon Y, the estimates for both the first and second moments, u_X and Σ_X, respectively, are dependent upon Y. Specifically, $S_W\{XY\}$ and $[S_W\{Y^2\}]^-$ are used for calculating all three of the estimates; furthermore, $\bar{w}\{Y\}$ is used in the estimate of μ_X.

2 Mixtures

For many problems, the type of distribution is not known. While it is simple to estimate the mean and covariance for a Gaussian distribution, the resulting distribution may not truly represent the distribution. Consider the data:

$$X = \{-15, -14, -14, -13, 7, 7, 8, 8, 8, 9, 9\}.$$

Here, $\bar{x} = 0$, $S_X = \frac{1218}{11}$, giving the estimated Gaussian $\mathcal{N}(0, \frac{1218}{11})$. However, this indicates that the expected value for this data is 0, but no value even near 0 occurs in the data. It appears, rather, that either this data comes from a totally different distribution or that it covers two different Gaussians. See [1, Figure 1.4].

While we would prefer to determine the single distribution from which this data comes, it is easier to cluster the data into different regions and to then say that each one comes from its own distribution. In this case, we can divide that data as:

$$X_1 = \{-15, -14, -14, -13\}$$

and

$$X_2 = \{7, 7, 8, 8, 8, 9, 9\},$$

which would then have more reasonable Gaussian estimates of $\mathcal{N}(-14, \frac{1}{2})$ and $\mathcal{N}(8, \frac{4}{7})$, respectively. There would, furthermore, be a distribution over the Gaussians themselves. That is, as four of the data points in X were considered to come from X_1 and seven were considered to come from X_2, the two separate Gaussians would get priors, or weights, of $\frac{4}{11}$ and $\frac{7}{11}$, respectively. So, for M Gaussians with respective means $\mu_1, \ldots, \mu_M$; covariances $\Sigma_1, \ldots, \Sigma_M$; and weights $w_1, \ldots, w_M$, where $\sum_{i=1}^{M} w_i = 1$, Gaussian mixtures are:

$$\sum_{i=1}^{M} w_i \mathcal{N}(\mu_i, \Sigma_i) \tag{14}$$

Similarly, we can have conditional Gaussian mixtures:

$$\sum_{i=1}^{M} w_i \mathcal{N}(u_i, \Sigma_i), \text{where } u_i = \mu_i + B_i\, y, \tag{15}$$

with a different regression matrix B_i as well for each mixture. Finally, let θ_{X_i} represent the parameters for mixture i of X, θ_X the parameters for all the mixtures of X, and θ_Y the parameters for Y.

3 Likelihoods

In this section I make the assumption that the elements of the covariance matrix for each mixture of X are zero off of its diagonal. In other words, the dimensions of the data X_i are uncorrelated with each other. Therefore, the covariance matrix Σ_{X_i} can be represented by its diagonal elements, contained in the variance vector $\sigma^2_{X_i}$. Put more simply, each dimension of X_i has its own one-dimensional Gaussian.

3.1 Gaussian Mixtures

3.1.1 Observed X

Given a data sample x from X, denote the P elements in the vector as $x[1], \ldots, x[P]$. Likewise, the elements of the mean vector and of the variance vector of $\mathcal{N}(\mu_i, \sigma_i^2)$, the Gaussian of mixture i, are denoted as $\mu_i[1], \ldots, \mu_i[P]$ and $\sigma_i^2[1], \ldots, \sigma_i^2[P]$, respectively. As each dimension is independent of each other, the likelihood is computed independently for each dimension:

$$\mathcal{L}(X[p] = x[p] | \theta_{X_i}) = \frac{\exp\left(\frac{-0.5(x[p]-\mu_i[p])^2}{\sigma_i^2[p]}\right)}{\sqrt{2\pi\sigma_i^2[p]}}. \tag{16}$$

The likelihoods for each dimension are then multiplied together:

$$\mathcal{L}(X = x | \theta_{X_i}) = \prod_{p=1}^{P} \mathcal{L}(X[p] = x[p] | \theta_{X_i}) \tag{17}$$

If there are mixtures of Gaussians, then we use a weighted sum of the likelihoods using (17) and w_i for each mixture i:

$$\mathcal{L}(X = x | \theta_X) = \sum_{i=1}^{M} w_i \mathcal{L}(X = x | \theta_{X_i}). \tag{18}$$

3.1.2 Partially-observed X

If any dimensions of x are missing, those dimensions need to be integrated out. Here we take advantage of the property of the Gaussian that, since it is a probability distribution, the integral over all of its domain is 1:

$$\int_{-\infty}^{\infty} \mathcal{L}(X[p]|\theta_{X_i})\, \mathrm{d}X[p] = 1, \tag{19}$$

hence, expanding (18), using (17), as:

$$\begin{aligned} \mathcal{L}(x|\theta_X) &= \sum_{i=1}^{M} w_i \prod_{p=1}^{P} f(x[p], \theta_{X_i}), \qquad (20) \\ \text{where } f(x[p], \theta_{X_i}) &= \begin{cases} 1 & , x[p] \text{ missing} \\ \mathcal{L}(X[p] = x[p]|\theta_{X_i}) & , x[p] \text{ observed} \end{cases} \end{aligned}$$

3.2 Conditional Gaussian mixtures

3.2.1 Observed Y

Given data samples x and y for a mixture of conditional Gaussians, we compute the likelihood as in (18) except that each mean μ_i is offset, or shifted, according to the value y and the regression weights B_i:

$$u_i = \mu_i + B_i\, y. \tag{21}$$

u_i is then substituted for μ_i in (16) to compute the conditional likelihood for a single dimension, single mixture, and all mixtures of a conditional Gaussian, respectively:

$$\mathcal{L}(X[p] = x[p]|Y = y, \theta_{X_i}, \theta_Y) = \frac{\exp\left(\frac{-0.5(x[p]-u_i[p])^2}{\sigma_i^2[p]}\right)}{\sqrt{2\pi\sigma_i^2[p]}} \tag{22}$$

$$\mathcal{L}(X = x|Y = y, \theta_{X_i}, \theta_Y) = \prod_{p=1}^{P} \mathcal{L}(X[p] = x[p]|Y = y, \theta_{X_i}, \theta_Y) \tag{23}$$

$$\mathcal{L}(X = x|Y = y, \theta_X, \theta_Y) = \sum_{i=1}^{M} w_i \mathcal{L}(X = x|Y = y, \theta_{X_i}, \theta_Y). \tag{24}$$

These are analogous to (16), (17), and (18), respectively. Thus, computing a likelihood with a conditional Gaussian with given data sample x and y is only slightly more complicated per mixture than having a regular Gaussian: there is an additional (matrix) multiplication and (vector) addition to compute each $u_i[p]$.

If any dimensions of x are missing in the conditional Gaussian, those dimensions can easily be integrated out. Regardless of whether y is observed or missing, we also have, similarly to (19):

$$\int_{-\infty}^{\infty} \mathcal{L}(X[p]|y, \theta_{X_i}, \theta_Y)\, \mathrm{d}X[p] = 1, \tag{25}$$

hence, expanding (24), in the case of observed y, as:

$$\begin{aligned} \mathcal{L}(x|Y = y, \theta_X, \theta_Y) &= \sum_{i=1}^{M} w_i \prod_{p=1}^{P} f(x[p], \theta_{X_i}, Y = y, \theta_Y), \qquad (26) \\ \text{where } f(x[p], \theta_{X_i}, Y = y, \theta_Y) &= \begin{cases} 1 & , x[p] \text{ missing} \\ \mathcal{L}(X[p] = x[p]|Y = y, \theta_{X_i}, \theta_Y) & , x[p] \text{ observed} \end{cases} \end{aligned}$$

3.2.2 Non-observed Y

However, suppose that we only have data sample x for a mixture of conditional Gaussians and that the sample y is missing. In other words, we do not know what the distribution is to be conditioned upon. The only items we have for y is its mean μ_Y and variance σ_Y^2, that is, its prior distribution. Let us, then, consider the likelihood for the first dimension of x in a given mixture i. Using this prior distribution, we can still shift its mean as follows (where $B_i[p]$ is row p of B_i):

$$u_i[1] = \mu_i[1] + B_i[1]\,\mu_Y \tag{27}$$

but need to also account for the variance of the unknown y in updating the variance of x, using the covariance between $X_i[1]$ and Y:

$$\begin{aligned} \sigma_{(X_i[1],Y)} &= B_i[1]\,\sigma_Y^2 && (28)\\ \hat{\sigma}_i^2[1] &= \sigma_i^2[1] + \sigma_{(X_i[1],Y)}\,B_i[1]. && (29) \end{aligned}$$

We then use the updated variance to calculate the likelihood of the first dimension ($p = 1$) for the given mixture i:

$$\mathcal{L}(X[1] = x[1]|Y = y, \theta_{X_i}, \theta_Y) = \frac{\exp\left[\frac{-0.5(x[1]-u_i[1])^2}{\hat{\sigma}_i^2[1]}\right]}{\sqrt{2\pi\hat{\sigma}_i^2[1]}}. \tag{30}$$

Having now seen the value for $x[1]$, the distribution for Y needs to be updated before using it to compute the likelihood for $x[2]$. That is, the difference between $x[1]$ and $\mu_i[1]$ will give an indication of where the hidden y may really be. For example, say that all elements of $B_i[1]$ are positive and that $x[1] > \mu_i[1]$; then we would expect y to be higher than the prior μ_Y. We, therefore, update the mean and variance for y as follows, letting $\hat{\theta}_Y = \{\hat{\mu}_Y, \hat{\sigma}_Y^2\}$:

$$\begin{aligned} \hat{\mu}_Y &= \mu_Y + \sigma_{(X_i[1],Y)}\,\frac{(x[1]-\mu_i[1])}{\hat{\sigma}_i^2[1]} && (31)\\ \hat{\sigma}_Y^2 &= \sigma_Y^2 - \sigma_{(X_i[1],Y)}\,\frac{\sigma_{(Y,X[1]),i}}{\hat{\sigma}_i^2[1]} && (32) \end{aligned}$$

Using (33), (34), (35), (36), (37), and (38), as follows, we then continue iteratively through the dimensions $2,\ldots,P$ of x for mixture i. We first compute three items, in a similar manner to (27), (28), and (29), respectively: $u_i[p]$; the updated covariance between $X_i[p]$ and Y; and the updated variance for Y:

$$\begin{aligned} u_i[p] &= \mu_i[p] + B_i[p]\,\hat{\mu}_Y && (33)\\ \hat{\sigma}_{(X_i[p],Y)} &= B_i[p]\,\hat{\sigma}_Y^2 && (34)\\ \hat{\sigma}_i^2[p] &= \sigma_i^2[p] + \hat{\sigma}_{(X_i[p],Y)}\,B_i[p]. && (35) \end{aligned}$$

We can then compute the likelihood for this dimension of x, as well as the updated distribution of Y, in a similar manner to (30), (31), and (32), respectively:

$$\begin{aligned} \mathcal{L}(X[p] = x[p]|Y = y, \theta_{X_i}, \hat{\theta}_Y) &= \frac{\exp\left(\frac{-0.5(x[p]-u_i[p])^2}{\hat{\sigma}_i^2[p]}\right)}{\sqrt{2\pi\hat{\sigma}_i^2[p]}} && (36)\\ \hat{\mu}_Y &= \hat{\mu}_Y + \hat{\sigma}_{(X_i[p],Y)}\,\frac{(x[p]-\mu_i[p])}{\hat{\sigma}_i^2[p]} && (37)\\ \hat{\sigma}_Y^2 &= \hat{\sigma}_Y^2 - \hat{\sigma}_{(X_i[p],Y)}\,\frac{\hat{\sigma}_{(Y,x[p]),i}}{\hat{\sigma}_i^2[p]}. && (38) \end{aligned}$$

The likelihood of the mixture is then

$$\mathcal{L}(X=x|Y=y,\theta_{X_i},\theta_Y)=\prod_{p=1}^{P}\mathcal{L}(X[p]=x[p]|Y=y,\theta_{X_i},\hat{\theta}_Y), \tag{39}$$

and the likelihood of all the mixtures is

$$\mathcal{L}(X=x|Y=y,\theta_X,\theta_Y)=\sum_{i=1}^{M}w_i\mathcal{L}(X=x|Y=y,\theta_{X_i},\theta_Y). \tag{40}$$

Note that for $p=1$, $\hat{\theta}_Y=\theta_Y$; in other words, the computation of the likelihood for each mixture starts with the prior for Y and then updates it according to x.

Hence, computing a likelihood with a conditional Gaussian with x given but y hidden involves a lot more computations due to inferring Y's updated distribution given dimensions $1,\ldots,P-1$ of the observed x (inferring the distribution after dimension P is not necessary for the likelihood computation).

If any elements of x are hidden in addition to y's being hidden, then (40) is expanded:

$$\begin{aligned}\mathcal{L}(X=x|Y,\theta_X,\theta_Y) &= \sum_{i=1}^{M}w_i\prod_{p=1}^{P}f(x[p],\theta_{X_i}|Y,\hat{\theta}_Y), \qquad (41)\\ \text{where } f(x[p],\theta_{X_i}|Y,\hat{\theta}_Y) &= \begin{cases}1 & ,x[p] \text{ missing}\\ \mathcal{L}(X[p]=x[p]|Y,\theta_{X_i},\hat{\theta}_Y) & ,x[p] \text{ observed}\end{cases}\end{aligned}$$

With hidden y, note that, if $x[p]$ is missing, the subsequent updates in (37) and (38) are not done in that iteration as $\hat{\mu}_Y$ and $\hat{\sigma}^2_Y$ are only updated for observed dimensions of x.

4 Conclusion

I myself have used conditional Gaussians in the context of Bayesian network based automatic speech recognition [2, 3]. In such a context, conditional Gaussians have been shown to be useful in conditioning the standard feature x upon an "auxiliary" feature y, thus enabling some of the correlation of the dimensions of X to be further modeled via the mutual conditioning upon Y. Furthermore, it is of potential benefit to use observations for y in estimating the parameters but to hide the observations for Y in calculating the likelihood.

For further information on Gaussians and their estimates, see [4] and [5, Chapter 2]. For further information on conditional Gaussians, see [6]. For further information on updating distributions using other random variables' observations, see [7].

Acknowledgments

This work was supported by the Swiss National Science Foundation under the grant BN_ASR (20-64172.00). Mathew Magimai-Doss was helpful in the preparation of this report.

References

[1] Kevin Patrick Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, University of California, Berkeley, 2002.

[2] Todd A. Stephenson, Jaume Escofet, Mathew Magimai-Doss, and Hervé Bourlard, "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables," in *Neural Networks for Signal Processing XII–Proceedings of the 2002 IEEE Signal Processing Society Workshop (NNSP 2002)*, Hervé Bourlard, Tülay Adali, Samy Bengio, Jan Larsen, and Scott Douglas, Eds., Martigny, Switzerland, September 2002.

[3] Todd A. Stephenson, Mathew Magimai-Doss, and Hervé Bourlard, "Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, April 2003, To appear.

[4] Athanasios Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., New York, third edition, 1991.

[5] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, Computer science and scientific computing. Academic Press, San Diego, second edition, 1990.

[6] Steffen L. Lauritzen and Frank Jensen, "Stable local computations with conditional Gaussian distributions," *Statistics and Computing*, vol. 11, no. 2, pp. 191–203, April 2001.

[7] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., 1988.