Fitting a Conditional Linear Gaussian Distribution

Kevin P. Murphy

28 October 1998
Revised 29 January 2003

1 Introduction

We consider the problem of finding the maximum likelihood (ML) estimates of the parameters of a conditional Gaussian variable Y with continuous parent X and discrete parent Q, i.e.,

$$p(y|x, Q = i) = c|\Sigma_i|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(y - B_i x - \mu_i)'\Sigma_i^{-1}(y - B_i x - \mu_i)\right)$$

where $c = (2\pi)^{-d/2}$ is a constant and $|y| = d$. The j'th row of B_i is the regression vector for the j'th component of y given that $Q = i$. We consider tying and various constraints on the covariance matrix in order to reduce the number of free parameters.

We will allow any of the variables to be hidden — we will replace observed values with expected values conditioned on evidence, as in EM. We express all the estimates in terms of expected sufficient statistics, whose size is independent of the number of samples. (This is different from the usual presentation, which give the formulas in terms of the raw data matrix.) The resulting formulas can be used in the M step of all of the following common models, which use special cases of the above equation:

- Factor analysis. Q does not exist, Σ is assumed diagonal, X is hidden and Y is observed. (The temporal version of this is the Kalman filter.)
- Mixture of Gaussians. X does not exist, Q is hidden, and Y is observed. (The temporal version of this is an HMM with MOG outputs.)
- Mixture of factor analyzers. Σ_i is diagonal, Q and X are hidden, Y is observed. (The temporal version of this is a switching Kalman filter.)

We assume that we have N i.i.d. training cases $\{e_t\}$, so the complete-data log-likelihood is

$$\log \prod_{t=1}^{N} \prod_{i=1}^{|Q|} [\Pr(y_t|x_t, Q_t = i, e_t)]^{q_t^i}$$

where $q_t^i = 1$ if Q has value i in the t'th complete case, and 0 otherwise. Since Q, X and Y may all be unobserved, we compute the expected complete-data log likelihood as follows (dropping terms which are independent of the parameters of Y)

$$L = -\tfrac{1}{2} \sum_t E\left[\sum_i q_t^i \log|\Sigma_i| + q_t^i (y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)|e_t\right]$$

By the chain rule, we can write

$$E[q_t^i x_t x_t'|e_t] = E[q_t^i|e_t]E[x_t x_t'|Q_t = i, e_t] \stackrel{\text{def}}{=} w_t^i E_{ti}[XX']$$

where the weights $w_t^i = \Pr(Q = i|e_t)$ are posterior probabilities (responsibilities), and $E_{ti}[XX']$ is a conditional second moment; we can rewrite the other moments similarly. In this new notation, the expected complete-data log-likelihood becomes

$$L = -\tfrac{1}{2}\sum_t \sum_i w_t^i \log|\Sigma_i| - \tfrac{1}{2}\sum_t \sum_i w_t^i E_{ti}\left[(y_t - B_i x_t - \mu_i)'\Sigma_i^{-1}(y_t - B_i x_t - \mu_i)\right] \quad (1)$$

To simplify future equations, we introduce the following expected sufficient statistics:

$$\begin{aligned}
w_i &\stackrel{\text{def}}{=} \sum_t w_t^i \\
S_{YY',i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[YY'] \\
S_{Y'Y,i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[Y'Y] \\
S_{Y,i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[Y] \\
S_{XX',i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[XX'] \\
S_{X,i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[X] \\
S_{XY',i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[XY'] \\
S_{YX',i} &\stackrel{\text{def}}{=} \sum_t w_t^i E_{ti}[YX']
\end{aligned}$$

Obviously $\sum_i w_i = N$, $S_{XY',i} = S_{XY',i}$, etc.

The goal is to derive the equations so we can implement a function of the form

$$(\mu_i, \Sigma_i, B_i) = \text{Mstep-clg}(w_i, S_{YY',i}, S_{Y'Y,i}, S_{Y,i}, S_{XX',i}, S_{X,i}, S_{XY',i})$$

For the no regression case, where X does not exist, we can simplify this to

$$(\mu_i, \Sigma_i) = \text{Mstep-cond-gauss}(w_i, S_{YY',i}, S_{Y'Y,i}, S_{Y,i})$$

2 Estimating the regression matrix

2.1 Untied

Using the following identity (see e.g., [Row99],[Jor03, ch.13])

$$\frac{\partial\left((Xa+b)'C(Xa+b)\right)}{\partial X} = (C + C')(Xa + b)a' \quad (2)$$

where $X = -B_i$, $a = x_t$, $b = y_t - \mu_i$, $C = \Sigma_i^{-1}$, we have

$$\begin{aligned}
\frac{\partial}{\partial B_i}L &= -\tfrac{1}{2}\sum_t w_t^i \cdot -2\Sigma_i^{-1} \cdot E_{ti}[(y_t - B_i x_t - \mu_i)x_t'] \\
&= \Sigma_i^{-1}\left\{(\sum_t w_t^i E_{ti}[YX']) - B_i(\sum_t w_t^i E_{ti}[XX']) - \mu_i(\sum_t w_t^i E_{ti}[X'])\right\} \\
&= \Sigma_i^{-1}\left\{S_{YX',i} - B_i S_{XX',i} - \mu_i S'_{X,i}\right\}
\end{aligned}$$

Setting $\frac{\partial}{\partial B_i} L = 0$ yields

$$\hat{B}_i = \left(S_{YX',i} - \mu_i S'_{X,i}\right) S^{-1}_{XX',i} \tag{3}$$

If we set $\mu_i = 0$, we recognize this as the (weighted) normal equations:

$$\hat{B}_i = S_{YX',i} S^{-1}_{XX',i}$$

2.2 Tied

The derivation is similar to the above.

$$\frac{\partial}{\partial B} L = \sum_i \Sigma_i^{-1} \left\{S_{YX',i} - B S_{XX',i} - \mu_i S'_{X,i}\right\}$$

Unfortunately, this is hard to solve. So we will assume the covariance is also tied, leading to

$$\frac{\partial}{\partial B} L = \Sigma^{-1} \sum_i \left\{S_{YX',i} - B S_{XX',i} - \mu_i S'_{X,i}\right\}$$

and hence

$$\hat{B} = \left(\sum_i S_{YX',i} - \sum_i \mu_i S'_{X,i}\right) \left(\sum_i S_{XX',i}\right)^{-1} \tag{4}$$

3 Estimating the mean

3.1 Untied

We can estimate μ_i similarly to B_i. Using Equation 2 where $X = \mu_i$, $a = 1$, $b = y_t - B_i x_t$, $C = \Sigma_i^{-1}$, we have

$$\begin{aligned}
\frac{\partial}{\partial \mu_i} L &= -\tfrac{1}{2} \sum_t w_t^i \cdot -2\Sigma_i^{-1} \cdot E_{ti}[(y_t - B_i x_t - \mu_i)] \\
&= \Sigma_i^{-1} \left\{ (\sum_t w_t^i E_{ti}[Y]) - B_i (\sum_t w_t^i E_{ti}[X]) - \mu_i (\sum_t w_t^i 1) \right\} \\
&= \Sigma_i^{-1} \left\{S_{Y,i} - B_i S_{X,i} - \mu_i w_i\right\}
\end{aligned}$$

Setting $\frac{\partial}{\partial \mu_i} L = 0$ yields

$$\hat{\mu}_i = \frac{S_{Y,i} - B_i S_{X,i}}{w_i} \tag{5}$$

3.1.1 No regression

If $B_i = 0$, this yields the familiar special case

$$\hat{\mu}_i = \frac{S_{Y,i}}{w_i} = \frac{\sum_t w_t^i E_{ti} Y}{\sum_t w_t^i} \tag{6}$$

3.2 Tied

$$\frac{\partial}{\partial \mu} L = \Sigma_i^{-1} \sum_i \left\{S_{Y,i} - B_i S_{X,i} - \mu_i w_i\right\}$$

So

$$\hat{\mu} = \frac{\sum_i (S_{Y,i} - B_i S_{X,i})}{N} \tag{7}$$

3.2.1 No regression

If $B = 0$, this yields

$$\hat{\mu} = \frac{\sum_i S_{Y,i}}{N} \tag{8}$$

4 Estimating the regression matrix and the mean simultaneously

Since the equation for B_i depends on μ_i and vice versa, if they are both to be estimated (as opposed to being clamped to fixed values), we must estimate them jointly. We can do this by appending μ_i as the last column to B_i to create A_i, and appending a 1 to the last component of X to create Z. Then the likelihood becomes

$$p(y|x, Q = i) = c|\Sigma_i|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(y - A_i z)'\Sigma_i^{-1}(y - A_i z)\right)$$

We use the equations from Section 2, with $\mu_i = 0$ and replacing $S_{XX',i}$ with $S_{ZZ',i}$ and $S_{YX',i}$ with $S_{ZY',i}$, defined below. Specifically,

$$\hat{A}_i = S_{YZ',i} S_{ZZ',i}^{-1} \tag{9}$$

The substitutions are

$$E_{ti} Z Z' = E_{ti} \begin{pmatrix} X \\ 1 \end{pmatrix} \begin{pmatrix} X' & 1 \end{pmatrix} = E_{ti} \begin{pmatrix} XX' & X \\ X' & 1 \end{pmatrix}$$

so

$$S_{ZZ',i} = \begin{pmatrix} S_{XX',i} & S_{X,i} \\ S'_{X,i} & w_i \end{pmatrix}$$

Also,

$$E_{ti} Z Y' = E_{ti} \begin{pmatrix} X \\ 1 \end{pmatrix} Y' = E_{ti} \begin{pmatrix} XY' \\ Y' \end{pmatrix}$$

so

$$S_{ZY',i} = \begin{pmatrix} S_{XY',i} \\ S'_{Y,i} \end{pmatrix}$$

5 Estimating a full covariance matrix

We assume the mean (whether estimated or clamped) is appended to A_i, and that a 1 is appended to Z, to simplify notation.

5.1 Untied

Using the identities

$$\frac{\partial \ln|X|}{\partial X} = (X')^{-1} \text{ and } \ln|X| = -\ln|X^{-1}|$$

we have

$$\frac{\partial}{\partial \Sigma_i^{-1}} \ln|\Sigma_i| = -\frac{\partial}{\partial \Sigma_i^{-1}} \ln|\Sigma_i^{-1}| = -\Sigma_i$$

Also, using the identity

$$\frac{\partial (a'Xb)}{\partial X} = ab'$$

where $a' = b = (y_t - A_i z_t)$ and $X = \Sigma_i^{-1}$, we have

$$\begin{aligned} \frac{\partial}{\partial \Sigma_i^{-1}} L &= \tfrac{1}{2}(\sum_t w_t^i)\Sigma_i - \tfrac{1}{2}\sum_t w_t^i E_{ti}(y_t - A_i z_t)(y_t - A_i z_t)' \\ &= 0 \end{aligned}$$

Hence

$$\begin{aligned}\hat{\Sigma}_i &= \frac{1}{w_i}\sum_t w_t^i E_{ti}\left(YY' - YZ'A_i' - A_iZY' + A_iZZ'A_i'\right) \\ &= \frac{1}{w_i}\left(S_{YY',i} - S_{YZ',i}A_i' - A_iS_{ZY',i} + A_iS_{ZZ',i}A_i'\right) \end{aligned} \tag{10}$$

If $A_i = \hat{A}_i$ in Equation 9, we have

$$A_iS_{ZZ',i}A_i' = \left(S_{YZ',i}S_{ZZ',i}^{-1}\right)S_{ZZ',i}\left(S_{ZZ',i}^{-1}S_{ZY',i}\right) = S_{YZ',i}A_i'$$

so the above simplifies further to

$$\hat{\Sigma}_i = \frac{1}{w_i}\left(S_{YY',i} - A_iS_{ZY',i}\right) \tag{11}$$

5.1.1 No regression

If $A_i = \mu_i$, Equation 10 simplifies to

$$\hat{\Sigma}_i = \frac{1}{w_i}\left(S_{YY',i} - S_{Y,i}\mu_i' - \mu_iS_{Y,i}' + \mu_i\mu_i'\right)$$

If in addition $\mu_i = \hat{\mu}_i = \frac{S_{Y,i}}{w_i}$ from Equation 6, then

$$\hat{\Sigma}_i = \frac{S_{YY',i}}{w_i} - \mu_i\mu_i' \tag{12}$$

5.2 Tied

We have

$$\frac{\partial}{\partial\Sigma^{-1}}L = \tfrac{1}{2}(\sum_i\sum_t w_t^i)\Sigma - \tfrac{1}{2}\sum_t\sum_i w_t^iE_{ti}(y_t - A_iz_t)(y_t - A_iz_t)'$$

Hence

$$\hat{\Sigma} = \frac{1}{N}\sum_i\left(S_{YY',i} - S_{YZ',i}A_i' - A_iS_{ZY',i} + A_iS_{ZZ',i}A_i'\right) \tag{13}$$

If $A_i = \hat{A}_i$, then

$$\hat{\Sigma}_i = \frac{1}{N}\sum_i\left(S_{YY',i} - A_iS_{ZY',i}\right) \tag{14}$$

5.2.1 No regression

If $A_i = \mu_i$, Equation 13 simplifies to

$$\hat{\Sigma} = \frac{1}{N}\sum_i\left(S_{YY',i} - S_{Y,i}\mu_i' - \mu_iS_{Y,i}' + \mu_i\mu_i'\right)$$

If in addition $\mu_i = \hat{\mu}_i = \frac{S_{Y,i}}{w_i}$ from Equation 6, then

$$\hat{\Sigma} = \frac{\sum_i S_{YY',i}}{N} - \sum_i\mu_i\mu_i' \tag{15}$$

6 Estimating a diagonal covariance matrix

Proceed as in estimating a full matrix, but then set all off-diagonal entries to 0.

7 Estimating a spherical covariance matrix

7.1 Untied

If we have the constraint that $\Sigma_i = \sigma_i^2 I$ is isotropic, the conditional density of Y becomes

$$p(y|x, Q = i) = c\sigma_i^{-d} \exp\left(-\tfrac{1}{2}\sigma_i^{-2}||y - A_i z||^2\right)$$

Hence

$$L = -d \sum_t \sum_i w_t^i E_{ti}[\log \sigma_i - \tfrac{1}{2}\sigma_i^{-2}||y - A_i z||^2]$$

so

$$\frac{\partial}{\partial \sigma_i} L = -d \sum_t w_t^i \sigma_i^{-1} + \sigma_i^{-3} w_t^i E_{ti}||y_t - A_i z_t||^2 = 0$$

and

$$\sigma_i^2 = \frac{1/d}{\sum_t w_t^i} \left(\sum_t w_t^i E_{ti}||y_t - A_i x_t||^2 \right)$$

Now

$$||y_t - A_i z_t||^2 = (y_t - A_i z_t)'(y_t - A_i z_t) = y_t' y_t + z_t' A_i' A_i z_t - 2 y_t' A_i z_t$$

To compute the expected value of this distance, we use the fact that $x'Ay = \text{tr}(x'Ay) = \text{tr}(Ayx')$, so $E[x'Ay] = \text{tr}(AE[yx'])$. Hence

$$\sum_t w_t^i E_{ti}(y_t - A_i z_t)'(y_t - A_i z_t) = \text{tr}(\sum_t w_t^i E_{ti} Y'Y) + \text{tr}(\sum_t w_t^i A_i' A_i E_{ti} ZZ') - 2\text{tr}(\sum_t w_t^i A_i E_{ti} ZY')$$

Now $\text{tr}(A) + \text{tr}(B) = \text{tr}(A + B)$, so

$$\sigma_i^2 = \frac{1/d}{w_i} \text{tr}\left(S_{Y'Y,i} + A_i' A_i S_{ZZ',i} - 2A_i S_{ZY',i}\right) \tag{16}$$

7.1.1 No regression

If $A_i = \mu_i$ and $z_t = 1$,

$$\sum_t w_t^i E_{ti}(y_t - A_i z_t)'(y_t - A_i z_t) = \sum_t w_t^i E_{ti} \left(Y'Y + \mu_i' \mu_i - 2Y'\mu_i\right)$$

If $\mu_i = \hat{\mu_i} = \frac{\sum_t E_{ti} Y}{w_i}$ as in Equation 5, this becomes

$$\sigma_i^2 = \frac{1}{d}\left(\frac{S_{Y'Y,i}}{w_i} - \mu_i' \mu_i\right) \tag{17}$$

7.2 Tied

If σ_i^2 is tied, we get

$$\sigma^2 = \frac{1/d}{N} \text{tr} \sum_i \left(S_{Y'Y,i} + A_i' A_i S_{ZZ',i} - 2A_i S_{ZY',i}\right) \tag{18}$$

7.2.1 No regression

For the tied case, we get

$$\sigma^2 = \frac{1}{Nd}\left(\sum_i S_{Y'Y,i} + \sum_i w_i \mu_i' \mu_i\right) \tag{19}$$

8 MAP estimates

You may encounter numerical problems when estimating CLG distributions, especially with small data sets or with mixture components that have low reponsibility (and hence little data assigned to them). A simple solution to this is to put a prior on the parameters, and compute maximum a posterior (MAP) estimates instead of maximum likelihood (ML) estimates.

Minka [Min00] discusses conjugate priors for the case of linear regression (including ridge regression, etc.) To extend these formulas to the current case, it would be necessary to derive the conditioning on Q, and to consider the partially observed case.

Most of the formulas needed for the no regression case have been derived in [HC95]; We summarize the results for the full-covariance, untied case below.[1] The tied and diagonal cases are similar. The details for the spherical case are not given, since regularization of a single scalar parameter is less important.

We put a Normal-Wishart prior on each Gaussian mixture component

$$P(\mu_i, \Sigma_i) = P(\mu_i)P(\Sigma_i|\mu_i) = \mathcal{N}(\mu_i; m_i, \tau_i^{-1} I_d) \times \mathcal{W}(\Sigma_i|\mu_i; \Lambda_i, \alpha_i)$$

where

$$\mathcal{W}(\Sigma_i|\mu_i; \Lambda_i, \alpha_i) \propto |\Sigma_i|^{(\alpha_i - d)/2} \exp(-\tfrac{1}{2}\mathrm{tr}(\Lambda_i \Sigma_i))$$

The mode of the Wishart is $\Sigma_i^{-1} = (\alpha_i - d)\Lambda_i^{-1}$, and the mean is $\Sigma_i^{-1} = \alpha_i \Lambda_i^{-1}$. Either of these can be used as initial estimates for Σ_i.

We can compute the MAP estimates by setting the derivative of the unnormalized log posterior to zero:

$$\begin{aligned}
\frac{\partial L^{MAP}}{\partial \mu_i} &= \left[\sum_t w_t^i E_{ti} \frac{\partial}{\partial \mu_i} \log \mathcal{N}(Y_t; \mu_i, \Sigma_i)\right] + \frac{\partial}{\partial \mu_i} \log \mathcal{N}(\mu_i; m_i, \tau_i) \\
&= \left[\sum_t w_t^i E_{ti} \Sigma_i^{-1}(Y_t - \mu_i)\right] - \tau_i \Sigma_i^{-1}(\mu_i - m_i)
\end{aligned}$$

so

$$\begin{aligned}
\hat{\mu}_i^{MAP} &= \frac{\tau_i m_i + \sum_t w_t^i E_{ti} Y_t}{\tau_i + \sum_t w_t^i} \\
&= \frac{\tau_i m_i + S_{Y,i}}{\tau_i + w_i}
\end{aligned} \tag{20}$$

Similarly,

$$\begin{aligned}
\frac{\partial L^{MAP}}{\partial \Sigma_i^{-1}} &= \left[\sum_t w_t^i E_{ti} \frac{\partial}{\partial \Sigma_i^{-1}} \log \mathcal{N}(Y_t; \mu_i, \Sigma_i)\right] + \frac{\partial}{\partial \Sigma_i^{-1}} \log \mathcal{W}(\Sigma_i) \\
&= \left[\sum_t w_t^i E_{ti} \tfrac{1}{2}\left(\Sigma_i - (Y_t - \mu_i)(Y_t - \mu_i)'\right)\right] + \left[\frac{\alpha_i - d}{2}\Sigma_i - \frac{\tau_i}{2}(\mu_i - m_i)(\mu_i - m_i)' - \tfrac{1}{2}\Lambda_i\right]
\end{aligned}$$

so

$$\begin{aligned}
\hat{\Sigma}_i^{MAP} &= \frac{\Lambda_i + \tau_i(\mu_i - m_i)(\mu_i - m_i)' + \sum_t w_t^i E_{ti}(Y_t - \mu_i)(Y_t - \mu_i)'}{\alpha_i - d + \sum_t w_t^i} \\
&= \frac{\Lambda_i + \tau_i(\mu_i - m_i)(\mu_i - m_i)' + S_{YY',i} - w_i \mu_i \mu_i'}{\alpha_i - d + w_i}
\end{aligned} \tag{21}$$

If we don't put a prior on μ_i (by setting the precision $\tau_i = 0$), this simplfies to

$$\hat{\Sigma}_i^{MAP} = \frac{\Lambda_i + S_{YY',i} - w_i \mu_i \mu_i'}{\alpha_i - d + w_i} \tag{22}$$

[1] We use slightly different notation. Specifically, we use μ_i as a parameter and m_i as a hyperparameter, whereas they use the opposite; we use the covariance matrix Σ_i instead of the precision matrix r_i^{-1}, and denote the prior covariance u_i by Λ_i. Note that τ_i is an inverse variance (scalar). Also, we use the expected value of Y.

A simple choice of hyper-parameters is $\Lambda_i = s_i I_d$ and $\alpha_i = d$, where s_i is some scaling factor, e.g., 0.01. (This can be implemented by simply replacing $S_{YY',i}$ with $S_{YY',i} + \Lambda_i$ in all the equations above.) Essentially this just regularizes the covariance estimate and avoids problems with singular matrices.

9 Deterministic annealing

EM is notorious for getting stuck in local optima. One approach is to use deterministic annealing [Ros98, UN98], slowly lowering a "temperature" parameter.

Brand [Bra99b, Bra99a] suggests optimizing

$$\theta^{MAP} = \arg\max_{\theta} \log P(D|\theta) - ZH(\theta)$$

where $-H(\theta)$ is a minimum entropy prior, $Z = T_0 - T$, and T is a temperature; he calls this "prior balancing". In the case of (unconditional) Gaussians, this becomes

$$\hat{\Sigma}^{MAP} = S_{YY'}/(N + Z)$$

Initially, T is a large positive number, so Z is a large negative number, which "inflates" $\hat{\Sigma}^{MAP}$. We reduce the temperature until $Z = 1$, the minimum entropy solution. ($Z = 0$ corresponds to maximum likelihood, and $Z = -1$ corresponds to maximum entropy.)

A similar approach can be applied to the more conventional Wishart prior: we start with s_i large, forcing all covariances to be broad, and hence all mixture components to receive a lot of support; then we gradually reduce the noise level.

10 Acknowledgments

Thanks to Rainer Deventer for careful proof-reading.

References

[Bra99a] M. Brand. Pattern discovery via entropy minimization. In *AI/Stats*, 1999. uncertainty99.microsoft.com/brand.htm.

[Bra99b] M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11:1155–1182, 1999.

[HC95] Qiang Huo and Chorkin Chan. Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 3(5S):334–345, 1995.

[Jor03] M. I. Jordan. An introduction to probabilistic graphical models, 2003. In preparation.

[Min00] T. Minka. Bayesian linear regression. Technical report, MIT, 2000.

[Ros98] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE*, 80:2210–2239, November 1998.

[Row99] S. Roweis. Matrix identities. Technical report, U. Toronto, 1999. www.cs.toronto.edu/ roweis/notes/matrixid.pdf.

[UN98] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.