

Dirichlet process mixture of linear models

Ewan Dunbar

July 10, 2011

In statistics, a *linear model* is one in which the data (the *response variable*) is understood to be generated by some distribution which can change in location—but only in location—in an additive way under the effect of some predictors. The term “linear model” by itself virtually always implies that the residual noise distribution is Gaussian. For example, a t-test implies a simple linear model, in which the predictor is a binary indicator variable marking which of the two groups the data point was drawn from. A slightly more sophisticated linear model is a regression, in which the predictor is continuous-valued. In general, there can be multiple predictors, and multiple response variables. In this general setting, a linear model is one in which the location of the Gaussian generating the data is set by multiplying some matrix of *coefficients* by some vector of *predictors*, as in (1):

(1)

$$\begin{aligned} \mathbf{y} &\sim N(\boldsymbol{\mu}, \Sigma) \\ \boldsymbol{\mu} &= A^T \mathbf{b} \end{aligned}$$

The coefficient matrix A must be of dimension $h \times d$, where h is one more than the number of predictors and d is the dimensionality of the data; if $A = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}$, for example, then the *intercept* (location for when all predictors are equal to zero) in the two-dimensional response space would be (β_{01}, β_{02}) , the effect of the first predictor would be to add (β_{11}, β_{12}) per unit value of the first predictor, so that the location in this case would be $(\beta_{01} + \beta_{11}, \beta_{02} + \beta_{12})$; and the effect of the second predictor would be to add (β_{21}, β_{22}) per unit value of the second predictor. For this to work out, \mathbf{b} needs to be the values of the predictor variables, augmented with a leading 1.

This ubiquitous model in statistics is also a simple model of a phonetic category which is affected by some rule. In particular, if in the perceptual system, phonetic rules are simple translations, then the correct model for a set of phonetic categories would be a mixture, not of Gaussians, but of linear models. For simplicity, assume that the vector of predictors is observed and is binary. It is important to note that linear models are subject to a *homogeneity of variance* assumption. That is, *only* the location changes as a function of the predictors, not the variance. Thus we will have the same constraint here.

Inference is straightforward. The standard base distribution for a Dirichlet process multivariate Gaussian mixture is a Normal Inverse Wishart distribution because the joint distribution of the two parameters of the Gaussian is then fully conjugate. In particular, the usual prior and posterior density are as in (2):

(2)

$$\begin{aligned} f(\boldsymbol{\mu}, \Sigma | m, \nu, \Psi, \kappa) &\propto |\Sigma|^{-\frac{d+\kappa+1}{2}} \nu^{\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\nu \Sigma^{-1} (\boldsymbol{\mu} - m)(\boldsymbol{\mu} - m)^T] \right\} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} \Psi] \right\} \\ f(\boldsymbol{\mu}, \Sigma | Y, m, \nu, \Psi, \kappa) &\propto |\Sigma|^{-\frac{d+\kappa^*+1}{2}} \nu^{*\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\nu^* \Sigma^{-1} (\boldsymbol{\mu} - m^*)(\boldsymbol{\mu} - m^*)^T] \right\} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1} \Psi^*] \right\} \\ m^* &= (N + \nu)^{-1} (N\bar{Y} + \nu m) \\ \nu^* &= \nu + N \\ \Psi^* &= \Psi + \sum_i^N (y_i - \bar{Y})(y_i - \bar{Y})^T + \frac{N\nu}{N + \nu} (\bar{Y} - m)(\bar{Y} - m)^T \\ \kappa^* &= \kappa + N \end{aligned}$$

However, in our case, there is no simple mean $\boldsymbol{\mu}$, but rather a matrix of regression coefficients A , which will be a simple mean only if there are no predictors. Not only the first row, but all the rows of A will be normal if A is a normally distributed matrix (Dawid 1981); that is, if A follows (3):

(3)

$$f(A|M, \Sigma, \Omega) \propto |\Sigma|^{-\frac{h}{2}} |\Omega|^{-\frac{d}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1}(A - M)^T \Omega^{-1}(A - M)] \right\}$$

A matrix normal distribution has the property that both its rows and its columns are normal. Σ is the column covariance matrix and Ω is the row covariance matrix. To make this clearer, note that, if A is matrix normally distributed, then $\text{vec}(A)$, the vectorization of A taken by going down the columns, is multivariate normal with mean $\text{vec}(M)$ and covariance $\Sigma \otimes \Omega$, so that the variance of the first element in the first column will be obtained by multiplying the first-column variance with the first-row variance, the variance of the second element in the first column will be obtained by multiplying the first-column variance with the second-row variance, their covariance will be obtained by multiplying the first-column variance with the first and second-row covariance, and so on. (This property also makes the distribution trivial to sample from.) For A of dimension $1 \times d$, the distribution reduces to a multivariate Gaussian with covariance $\omega\Sigma$. Substituting the matrix normal distribution in for the normal portion of the Normal Inverse Wishart, and substituting in a normal likelihood centered at $A^T B$, we get a conjugate prior as in (4) (assume the data Y is a $d \times N$ matrix of observations):

(4)

$$\begin{aligned} f(A, \Sigma|M, \Omega, \Psi, \kappa) &\propto |\Sigma|^{-\frac{h}{2}} |\Omega|^{-\frac{d}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1}(A - M)^T \Omega^{-1}(A - M)] \right\} \\ &\times |\Sigma|^{-\frac{d+\kappa+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1}\Psi] \right\} \\ f(A, \Sigma|Y, M, \Omega, \Psi, \kappa) &\propto |\Sigma|^{-\frac{h}{2}} |\Omega^*|^{-\frac{d}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1}(A - M^*)^T \Omega^{*-1}(A - M^*)] \right\} \\ &\times |\Sigma|^{-\frac{d+\kappa^*+1}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\Sigma^{-1}\Psi^*] \right\} \end{aligned}$$

$$M^* = \Omega^* E^T$$

$$\Omega^* = (\Omega^{-1} + B B^T)^{-1}$$

$$\Psi^* = Y Y^T + M^T \Omega^{-1} M - E \Omega^* E^T$$

$$\kappa^* = \kappa + N$$

where

$$E = X B^T + M^T \Omega^{-1}$$

with normalizing constant

$$K = \frac{|\Psi^*|^{\frac{\kappa^*}{2}}}{(2\pi)^{\frac{dh}{2}} (2^{\frac{d\kappa^*}{2}}) \Gamma_d(\frac{\kappa^*}{2}) |\Omega^*|^{\frac{d}{2}}}$$

This is sufficient detail to implement a Dirichlet process mixture; parameters can be resampled straightforwardly, a hyperprior can be put on the base distribution parameters (M , for example, with a matrix normal hyperprior and Ω with an Inverse Wishart), and on α (see West 1992, Neal 1998).

References

- DAWID, A.P. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* 68.265.
- NEAL, RADFORD M. 1998. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Technical Report 2.
- WEST, MIKE, 1992. Hyperparameter estimation in Dirichlet process mixture models.