Statistical Techniques in Robotics (16-831, F08) Lecture #14 (October 9)

Bayesian Linear Regression Pt. 2, Gaussian Properties

Lecturer: Drew Bagnell *Scribe: Hans Pirnay*

1 Parameterizations for Gaussians

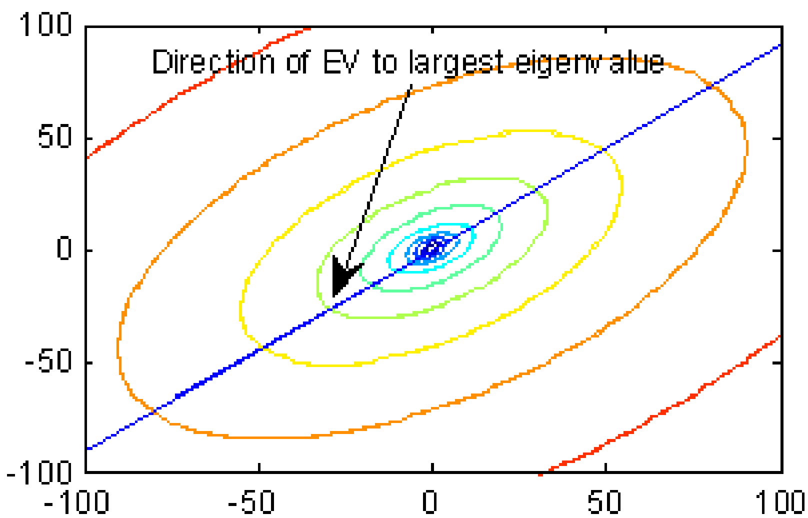
There are two common parameterizations for Gaussians, the moment parameterization and the natural parameterization.

The Moment Parameterization has the form

$$\mathcal{N}(\mu, \Sigma) = p(\theta) = \frac{1}{z} \exp\left(-\frac{1}{2}(\theta-\mu)\Sigma^{-1}(\theta-\mu)\right) \quad (1)$$

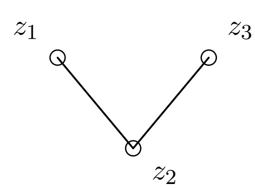
The Natural Parameterization is

$$\tilde{\mathcal{N}}(J, P) = \tilde{p}(\theta) = \frac{1}{z} \exp\left(J^T\theta - \frac{1}{2}\theta^T P\theta\right) \quad (2)$$

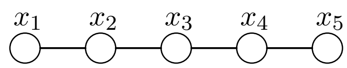


The matrix P of the natural parameterization has a graphical model interpretation. If there is a non-zero entry for (z_1, z_2), then there is a correspondence.

$$P = \boxed{\begin{matrix} X & X & 0 \\ X & X & X \\ 0 & X & X \end{matrix}}$$



Following the graphical model interpretation, P is in many cases highly structured. Consider for example the graphical model of a markov chain



This corresponds to a band structure in P:

$$P = \begin{pmatrix} X & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 \\ 0 & X & X & X & 0 \\ 0 & 0 & X & X & X \\ 0 & 0 & 0 & X & X \end{pmatrix} \tag{3}$$

Note: P^{-1} is, in general, not sparse! (this makes intuitve sense since $P^{-1} = \Sigma$ the covariance matrix, and the covariance of two states along the markov chain are not independent.)

2 Bayes Linear Regression Update

Scalar version of the likelihood field:

$$p(y|x,\theta) = \mathcal{N}(\theta^T x_t, \sigma_t^2) = \frac{1}{z} \exp\left(\frac{-(\theta^T x - y)(\theta^T x - y)}{2\sigma^2} \right) \tag{4}$$

(Don't worry about the weird notation of $\mathcal{N}$ as a function of σ^2. This is an arbitrary definition)

2.1 Deriving the update rules

Apply Bayes' Rule to the probability of a weight vector θ given a datapoint D.

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{z} \tag{5}$$

This results in the multiplication of two exponential functions. Adding the exponent of the prior to that of the likelihood yields

$$-\frac{1}{2\sigma^2}\left(\theta^T x - y\right)^2 + J^T\theta - \frac{1}{2}\theta^T P\theta \tag{6}$$

collecting terms to find updates J'_θ and P'_θ:

$$= -\frac{1}{2\sigma^2}\left(\theta^T x x^T \theta - 2\theta^T x y + y^2\right) + J^T\theta - \frac{1}{2}\theta^T P\theta \tag{7}$$

$$= \left(\frac{x^T y}{\sigma^2} + J^T\right)\theta - \frac{1}{2}\theta^T\left(\frac{x x^T}{\sigma^2} + P\right)\theta - \frac{y^2}{2\sigma^2} \tag{8}$$

Since this all happens in the exponent of an exponential function, the constat y^2-term can be shifted into the regularizing z. Thus, the update rules for J'_θ and P'_θ are

$$J'_\theta = \frac{xy}{\sigma^2} + J \tag{9}$$
$$P'_\theta = \frac{xx^T}{\sigma^2} + P \tag{10}$$

1. in a gaussian model, a new datapoint always lowers the variance - this downgrading of the variance does not always make sense
2. if you believe there are outliers, this model won't work for you
3. the variance is not a function of y. The precision if only affected by input not output. This is a consequence of having the same σ (observation error) everywhere in space.

2.2 Transfer to moment parameterization

The update rules for the natural parameterization at timestep t are

$$J \mathrel{+}= \frac{y_t x_t}{\sigma^2} \tag{11}$$
$$P \mathrel{+}= \frac{1}{\sigma^2} x x^T. \tag{12}$$

Having no prior knowledge about the data, we choose standard initial conditions

$$J_0 = 0 \tag{13}$$
$$P_0 = \mathbb{I}, \tag{14}$$

$\mathbb{I}$ being the identity matrix. Given the transfer rules to the moment parameterization

$$\Sigma = P^{-1} \tag{15}$$
$$\mu = P^{-1} J \tag{16}$$

the moment parameterization after N timesteps is then

$$\Sigma_\theta = \left[\sum_{i=1}^{N} \frac{x_i x_i^T}{\sigma^2} + \mathbb{I} \right]^{-1} \tag{17}$$
$$\mu_\theta = \left(\sum_{i=1}^{N} \frac{x_i x_i^T}{\sigma^2} + \mathbb{I} \right) \sum_{t=1}^{N} \frac{y_t x_t}{\sigma^2} \tag{18}$$

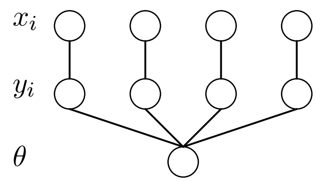
1. $\sum_{t=1}^{N} \frac{y_t x_t}{\sigma^2}$ is the gradient of Bayes online linear regression
2. this looks just like Newton's method
3. Computation time: $o(d^2)$ for update, $o(d^3)$ for mean (that can be reduced to $o(d^2)$ with tricks)

2.3 Making predictions

Given all data D up to timestep t and x_{t+1}, the probability of an observation $\tilde{y}_{t+1}$ is

$$\begin{aligned} p(\tilde{y}_{t+1}|x_{t+1}, D) &= \int p(\tilde{y}_{t+1}|x_{t+1}, D, \theta) \cdot p(\theta|D) d\theta & (19) \\ &= \int p(\tilde{y}_{t+1}|x_{t+1}, \theta) \cdot p(\theta, D) d\theta & (20) \end{aligned}$$

To know $p(\tilde{y}_{t+1}|x_{t+1}, D)$, we only need y and σ^2, because these parameters determine the gaussian.



2.4 Marginal and Conditional Distributions in different parameters

These computations are crucial for gaussian processes and Kalman filters.

2.4.1 Moment parameterization

Given:

$$\mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{21}$$

Marginal: computing $p(x_2)$

$$\begin{aligned} \mu_2^{\text{marg}} &= \mu_2 & (22) \\ \Sigma_2^{\text{marg}} &= \Sigma_{22} & (23) \end{aligned}$$

Conditional: computing $p(x_1|x_2)$

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) & (24) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & (25) \end{aligned}$$

2.4.2 Natural parameterization

Given:

$$\mathcal{N}\left(\begin{bmatrix} J_1 \\ J_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}\right) \tag{26}$$

Marginal: computing $p(x_2)$

$$J_2^{\mathrm{marg}} = J_2 - P_{21}P_{11}^{-1}J_1 \tag{27}$$

$$P_1^{\mathrm{marg}} = P_{12} - P_{21}P_{11}^{-1}P_{12} \tag{28}$$

Conditional: computing $p(x_1|x_2)$

$$J_{1|2} = J_1 - P_{12}x_2 \tag{29}$$

$$P_{1|2} = P_{11} \tag{30}$$