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

Bayesian Linear Regression Pt. 2, Gaussian Properties

Lecturer: Drew Bagnell *Scribe: Laura Lindzey (2009) Hans Pirnay(2008)*

1 Parameterizations for Gaussians

There are two common parameterizations for Gaussians, the moment parameterization and the natural parameterization. It is often most practical to switch back and forth between representations, depending on which calculations are needed.

1.1 Moment Parameterization

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

Given:

$$\mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \tag{2}$$

Marginal: computing $p(x_2)$

$$\mu_2^{\text{marg}} = \mu_2 \tag{3}$$
$$\Sigma_2^{\text{marg}} = \Sigma_{22} \tag{4}$$

Conditional: computing $p(x_1|x_2)$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1} (x_2 - \mu_2) \tag{5}$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{6}$$

1.2 Natural Parameterization

$$\tilde{\mathcal{N}}(J, P) = \tilde{p}(\theta) \quad = \quad \frac{1}{z} \exp\left(J^T\theta - \frac{1}{2}\theta^T P \theta \right) \tag{7}$$

Given:

$$\mathcal{N}\left(\begin{bmatrix} J_1 \\ J_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \right) \tag{8}$$

Marginal: computing $p(x_2)$

$$J_2^{\text{marg}} = J_2 - P_{21}P_{11}^{-1}J_1 \tag{9}$$
$$P_1^{\text{marg}} = P_{12} - P_{21}P_{11}^{-1}P_{12} \tag{10}$$

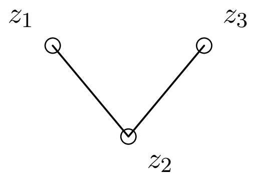
Conditional: computing $p(x_1|x_2)$

$$\begin{aligned} J_{1|2} &= J_1 - P_{12}x_2 && (11) \\ P_{1|2} &= P_{11} && (12) \end{aligned}$$

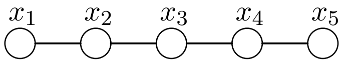
2 Graphical Models

The matrix P of the natural parameterization has a graphical model interpretation. If and only if there is a non-zero entry for (z_i, z_j), then there is a lik between z_i and z_j in the MRF that corresponds to $\tilde{N}(J, P)$.

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



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



The corresponding matrix P will be non-zero only on the diagonal and off-diagonal:

$$P = \begin{pmatrix} X & X & 0 & 0 & 0 \\ X & X & X & 0 & 0 \\ 0 & X & X & X & 0 \\ 0 & 0 & X & X & X \\ 0 & 0 & 0 & X & X \end{pmatrix} \qquad (13)$$

Note: P describes which variables **directly** affect each other.

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.)

3 Update Rules for Bayes Linear Regression

We want to derive how to update our representation given a series of new data $D = \{x_1, y_1, ..., x_n, y_n\}$. Let the prior be θ, where $\theta \sim N(\mu_0, \Sigma_0)$. Then, we want to calculate

$$p(y_t|D) \quad = \int p(y_t|\theta, x_t)p(\theta|D_t) \quad = E_{p(\theta|D)}[p(y_t|\theta, x] \qquad (14)$$

So, we need to calculate $p(\theta|D)$:

$$p(y|x,\theta) \;=\; \mathcal{N}(\theta^T x_t, \sigma_t^2) = \frac{1}{z}\exp\left(\frac{-(\theta^T x - y)(\theta^T x - y)}{2\sigma^2}\right) \tag{15}$$

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

3.1 Deriving the update rules

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

$$p(\theta|D) \;=\; \frac{p(D|\theta)p(\theta)}{z} \tag{16}$$

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

$$-\frac{1}{2\sigma^2}\left(\theta^T x - y\right)^2 + J^T\theta - \frac{1}{2}\theta^T P\theta \tag{17}$$

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

$$\begin{aligned}
&= -\frac{1}{2\sigma^2}\left(\theta^T x x^T\theta - 2\theta^T xy + y^2\right) + J^T\theta - \frac{1}{2}\theta^T P\theta && (18)\\
&= \left(\frac{x^T y}{\sigma^2} + J^T\right)\theta - \frac{1}{2}\theta^T\left(\frac{xx^T}{\sigma^2} + P\right)\theta - \frac{y^2}{2\sigma^2} && (19)
\end{aligned}$$

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

$$\begin{aligned}
J'_\theta &= \frac{xy}{\sigma^2} + J && (20)\\
P'_\theta &= \frac{xx^T}{\sigma^2} + P && (21)
\end{aligned}$$

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.

3.2 Transfer to moment parameterization

The update rules for the natural parameterization at timestep t are

$$\begin{aligned}
J &\mathrel{+}= \frac{y_t x_t}{\sigma^2} && (22)\\
P &\mathrel{+}= \frac{1}{\sigma^2}xx^T. && (23)
\end{aligned}$$

Given the transfer rules to the moment parameterization

$$\Sigma = P^{-1} \tag{24}$$

$$\mu = P^{-1}J \tag{25}$$

the moment parameterization after N timesteps is then

$$\Sigma_{\theta|D} = \left[\Sigma_0^{-1} + \sum_{i=1}^{N} \frac{x_i x_i^T}{\sigma^2}\right]^{-1} \tag{26}$$

$$\mu_\theta = \left[\Sigma_0^{-1} + \sum_{i=1}^{N} \frac{x_i x_i^T}{\sigma^2}\right] \sum_{t=1}^{N} \frac{y_t x_t}{\sigma^2} \tag{27}$$

1. If $\mu_0 \neq 0$, the update for μ_θ will be more complicated
2. $\sum_{t=1}^{N} \frac{y_t x_t}{\sigma^2}$ is the gradient of Bayes online linear regression
3. this looks just like Newton's method
4. Computation time: $o(d^2)$ for update, $o(d^3)$ for mean (that can be reduced to $o(d^2)$ with tricks)

3.3 Making predictions

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

$$p(\tilde{y}_{t+1}|x_{t+1}, D) = \int p(\tilde{y}_{t+1}|x_{t+1}, D, \theta) \cdot p(\theta|D) d\theta \tag{28}$$

$$= \int p(\tilde{y}_{t+1}|x_{t+1}, \theta) \cdot p(\theta, D) d\theta \tag{29}$$

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