

PRML

Chapter 3

2018.10.4

3.3 베이지안 선형 회귀

3.3.1 Parameter 분포

3.3.2 예측분포

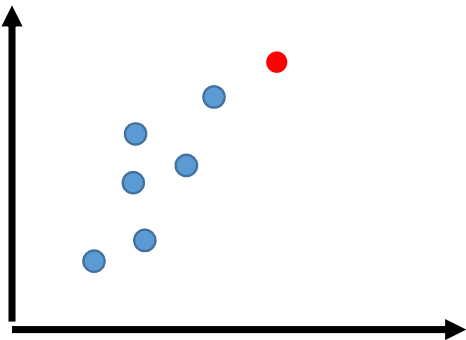
3.3.2 등가커널

3.4 베이지안 모델 비교

혹시, 공액 사전분포(Conjugate prior distribution) 기억 나시나요?

- Prior의 분포와 동일한 혹은 연관이 있는 분포로 Posterior를 만들고 싶을 때 사용하는 분포가 공액 사전분포 입니다.
- MLE 추정 방식과 다르게 데이터를 전부 사용 하지 않고 추정이 가능하며(Online 학습)
- Parameter w 를 점 추정하지 않고, 분포를 추정하여 Overfitting에 대해 좀더 강건한 것 같습니다.

뇌피셜..



붉은 색 데이터가 새로운 Input이 된다면..

1. MLE : $(X^T X)^{-1} X Y$
2. Bayes: 이미 추정한 w 에 업데이트 가능

3.3 Bayesian Linear Regression

- 베이زي안 방식으로 회귀 분석을 하는 방법을 알아 보겠습니다.

$$p(w) \sim N(w|m_0, S_0) \leftarrow \text{사전 확률 분포}$$

$$p(w|t) \sim N(w|m_N, S_N) \leftarrow \text{사후 확률 분포}$$

사후 확률 분포가 정규분포를 따르므로 $w_{MAP} = m_N$

- 계산의 편의를 위해 사전분포를 간단하게 바꿔보겠습니다.

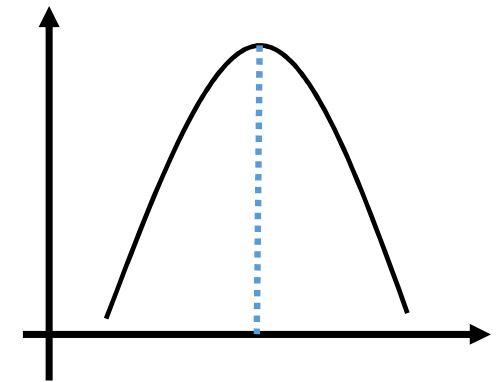
$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

이 때의 사후 분포의 파라미터는 다음과 같이 된다

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^T \Phi$$

기저함수

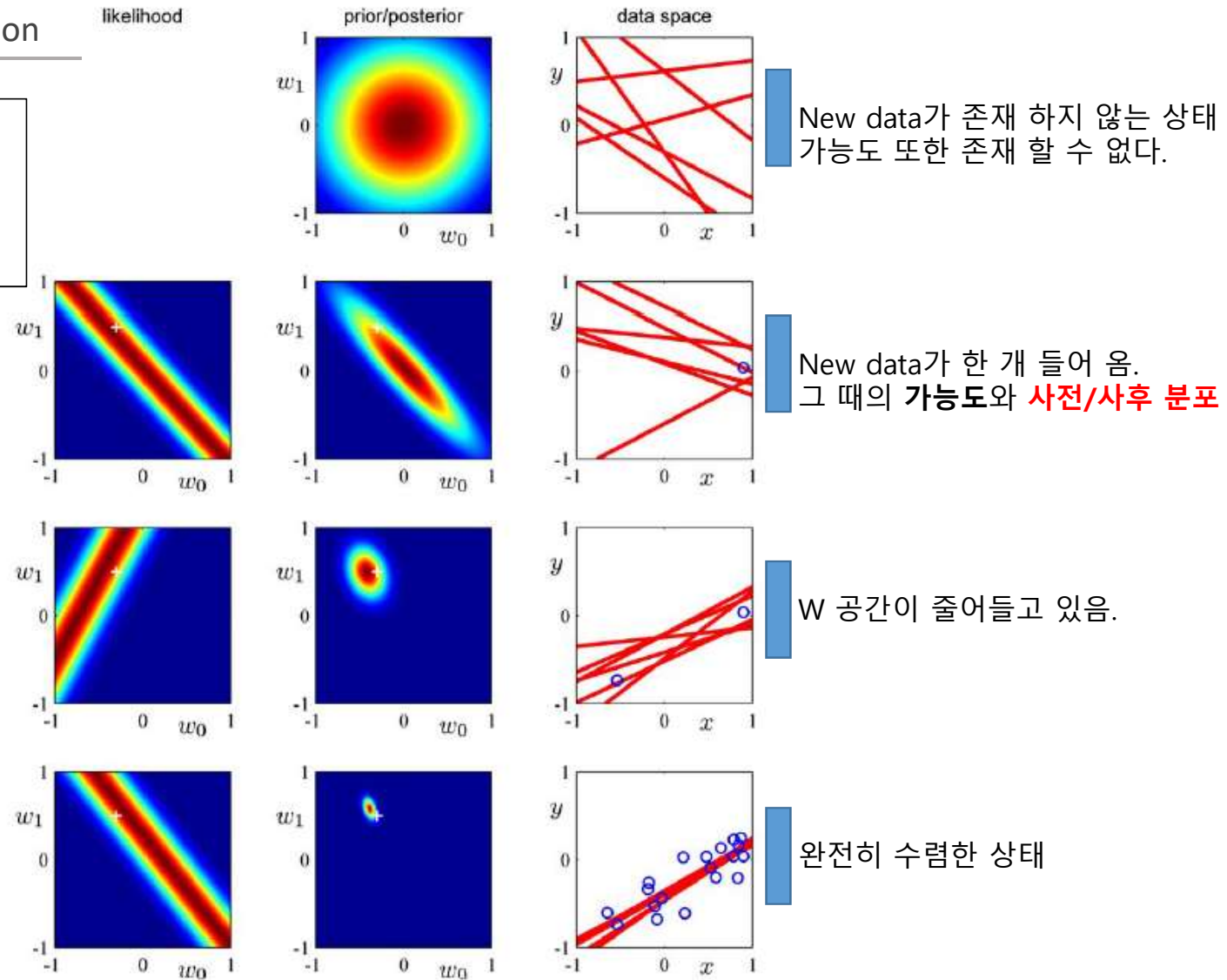


아무리, 베이زي안 회귀분석이
새로운 데이터를 통해
W를 Update를 해도 결국
사전 분포의 한계?를 벗어나지
못 하는 형태

3.3 Bayesian Linear Regression

추정하는 회귀식 $y = 0.5x - 0.3$

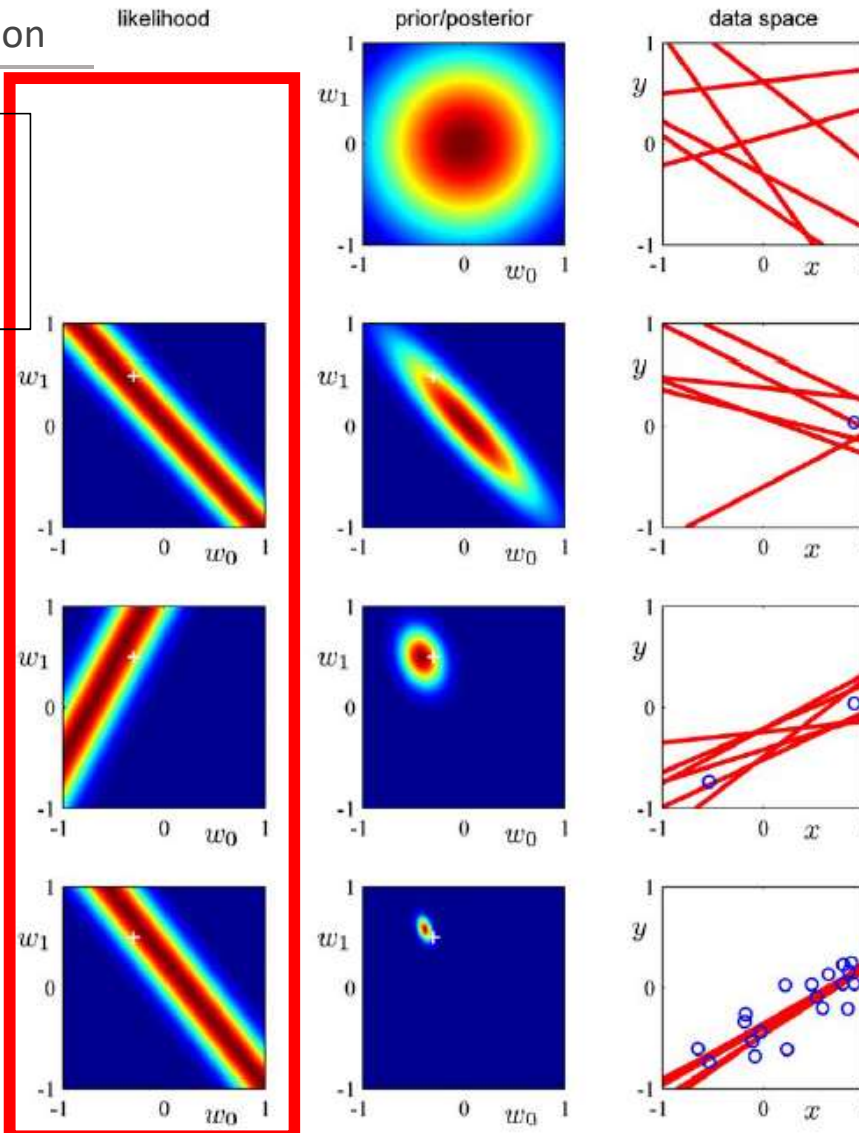
- 단계별 모델을 5개씩 만들
- W공간이 점점 줄어들고 있음



3.3 Bayesian Linear Regression

추정하는 회귀식 $y = 0.5x - 0.3$

- 단계별 모델을 5개씩 만들
- W공간이 점점 줄어들고 있음



New data가 존재 하지 않는 상태
가능도 또한 존재 할 수 없다.

New data가 한 개 들어 옴.
그 때의 **가능도**와 **사전/사후 분포**

W 공간이 줄어들고 있음.

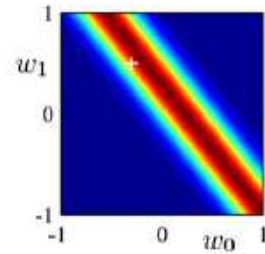
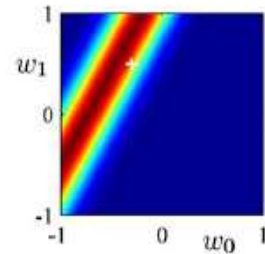
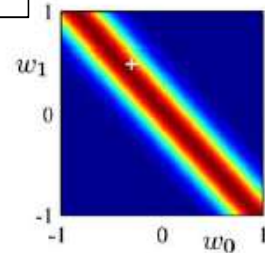
완전히 수렴한 상태

3.3 Bayesian Linear Regression

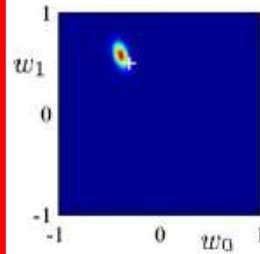
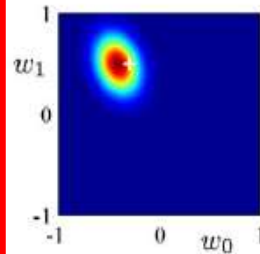
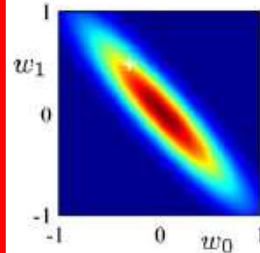
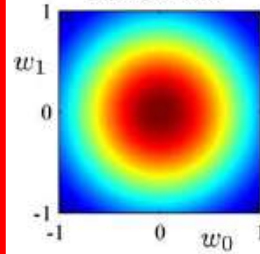
추정하는 회귀식 $y = 0.5x - 0.3$

- 단계별 모델을 5개씩 만들
- W공간이 점점 줄어들고 있음

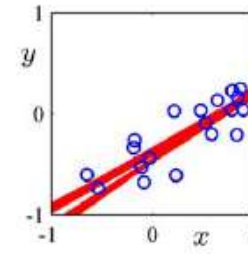
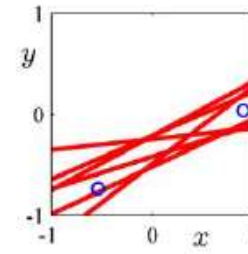
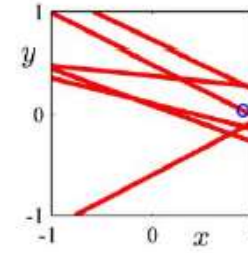
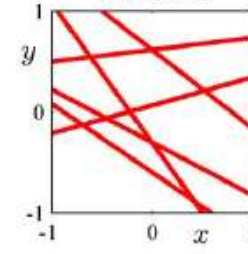
likelihood



prior/posterior



data space



New data가 존재 하지 않는 상태
가능도 또한 존재 할 수 없다.

New data가 한 개 들어 옴.
그 때의 **가능도**와 **사전/사후 분포**

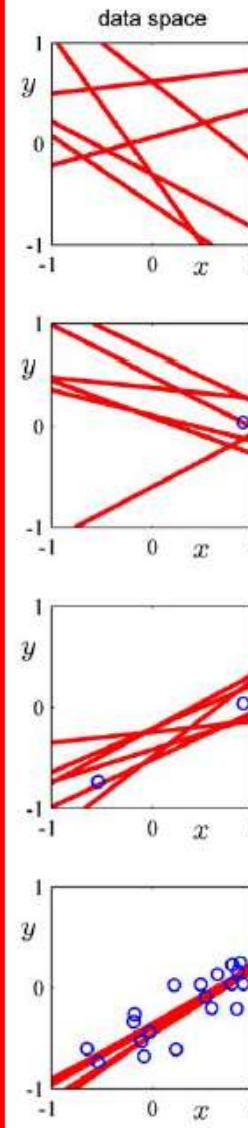
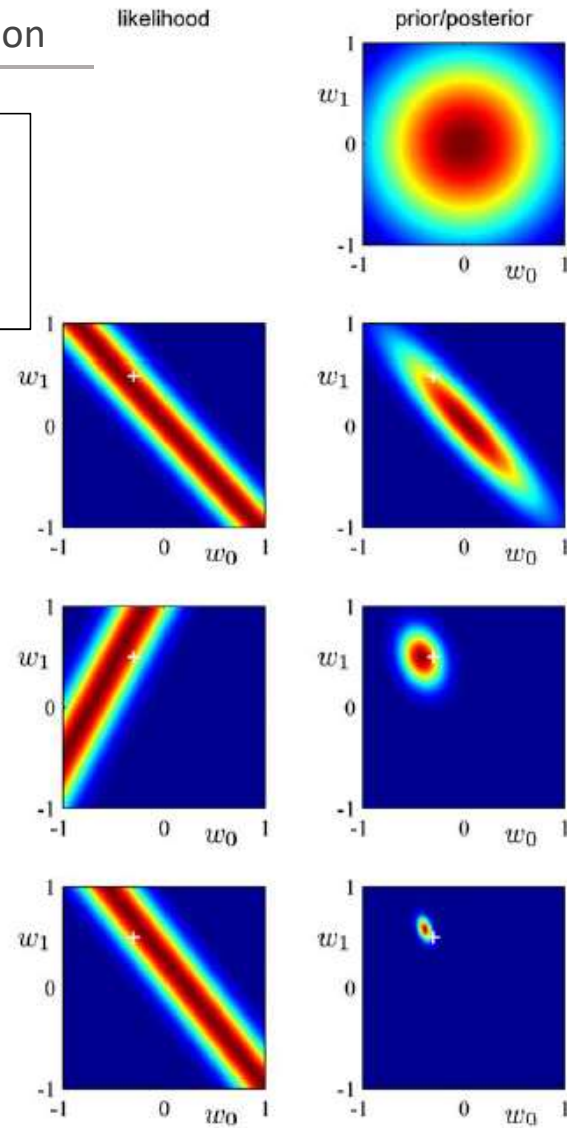
W 공간이 줄어들고 있음.

완전히 수렴한 상태

3.3 Bayesian Linear Regression

추정하는 회귀식 $y = 0.5x - 0.3$

- 단계별 모델을 5개씩 만들
- W공간이 점점 줄어들고 있음



New data가 존재 하지 않는 상태
가능도 또한 존재 할 수 없다.

New data가 한 개 들어 옴.
그 때의 **가능도**와 **사전/사후 분포**

W 공간이 줄어들고 있음.

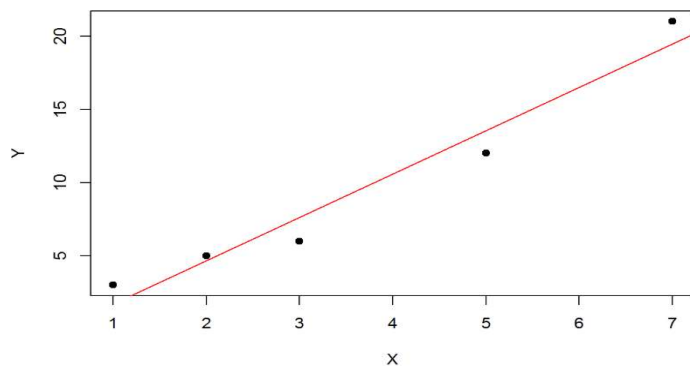
완전히 수렴한 상태

기저함수 예시

```
one <- function(x) rep(1,length(x))
id <- function(x) x
sq <- function(x) x^2
x3 <- function(x) x^3
x4 <- function(x) x^4
```

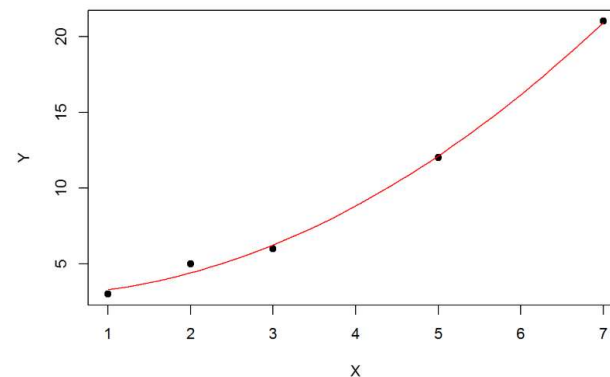
some data

```
X <- c(1,2,3,5,7)
Y <- c(3,5,6,12,21)
# basis for linear regression
phi <- c(one, id)
W <- compute_w(X, Y, phi)
plot(X,Y,pch=19)
abline(W, col="red")
```



basis for quadratic regression

```
phi <- c(one, id, sq)
W <- compute_w(X, Y, phi)
plot(X,Y,pch=19)
draw_regression(X,W,phi)
```



선형 회귀에 맞는 기저함수를 이용한다.

```
compute_posterior <- function(X, Y, m_old, S_old, phi= c(one, id)) {  
  Phi <- sapply(phi, function(base) base(X)) # make design matrix  
  if(length(X)==1) # 길이가 1이면, 벡터로 반환해서 matrix로 변형!  
  Phi <- t(as.matrix(Phi))  
  S_new <- solve(solve(S_old) + beta * t(Phi) %*% Phi)  
  m_new <- S_new %*% (solve(S_old) %*% m_old + beta * t(Phi) %*% Y)  
  list(m=m_new, S=S_new) # Update된 Parameter 반환  
}
```

$$m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T y)$$

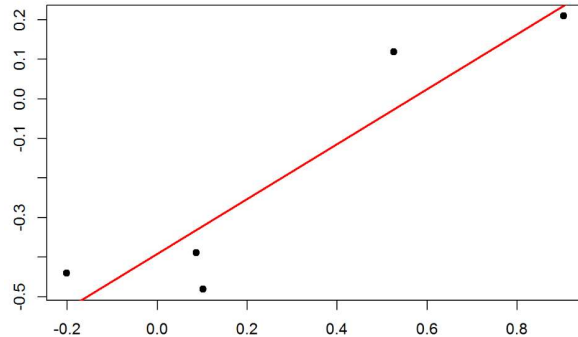
$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$$

```
alpha <- 2.0
m_0 <- c(0,0) # 평균을 0으로 맞춰준다.
S_0 <- alpha*diag(2) # 공분산 행렬 정의

set.seed(121)

X <- make.X(5) # 5개의 데이터 포인트를 이용하면!
Y <- make.Y(X)

posterior_1 <- compute_posterior(X, Y, m_0, S_0)
posterior_1$m
```

$$\begin{bmatrix} 0.391059 \\ 0.693193 \end{bmatrix}$$


```
plot(X, Y, pch=19, col="black")
abline(posterior_1$m, col="red", lwd=2)
```

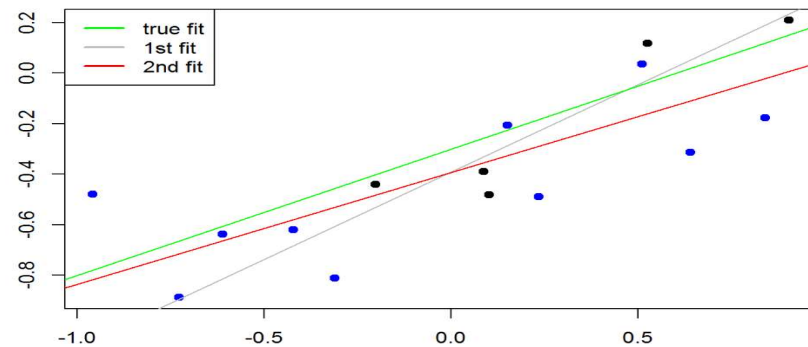
새로운 데이터 추가!

```
X_new <- make.X(10) # more points are available!
Y_new <- make.Y(X_new)
```

```
posterior_2 <- compute_posterior(X_new, Y_new, posterior_1$m,
posterior_1$S)
posterior_2$m
```

$$\begin{bmatrix} -0.3930011 \\ 0.4430177 \end{bmatrix}$$

```
plot(c(X,X_new),c(Y,Y_new),type="n")
legend("topleft",c("true fit","1st fit","2nd fit"),
col=c("green","grey","red"), lty=1, lwd=2)
points(X , Y , pch=19, col="black")
points(X_new, Y_new, pch=19, col="blue")
abline(posterior_1$m, col="grey") # old fit
abline(posterior_2$m, col="red") # new fit
abline(c(-0.3,.5), col="green")
```



이제껏, W 추정에 힘을 쓰신분 계신가요,,?

- 실제 응용 사례에서는 W 값을 알아내는 것 보다는 새로운 값에 대하여 t 의 값을 예측하는 것이 더 중요하다.
- 다음과 같이 정의 되는 **예측분포(Predictive distribution)**를 보자!

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w}$$

- 사용 하는 변수

Train set으로 부터 주어진 target 벡터 : \mathbf{t}

그에 대한 출력 (예측하려는) : t

모델 중 매개 변수 : W

Precision : β

정규분포 일반화 계수: α

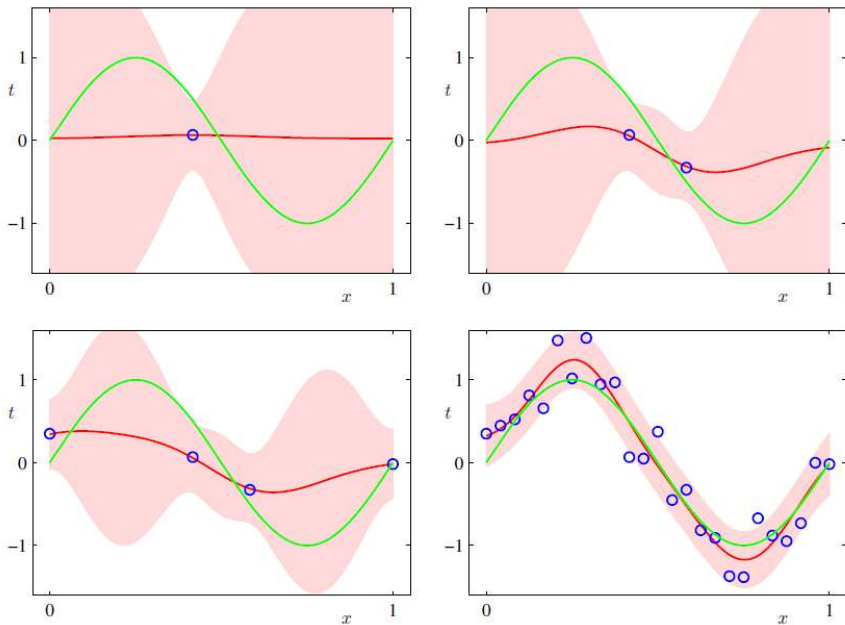
3.3 Bayesian Linear Regression - Predictive distribution

- 앞선 수식들을 잘 조합한다면 (해보지 않았지만...) 아래와 같이 정리 됩니다.

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \phi(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

where the variance $\sigma_N^2(\mathbf{x})$ of the predictive distribution is given by

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}).$$



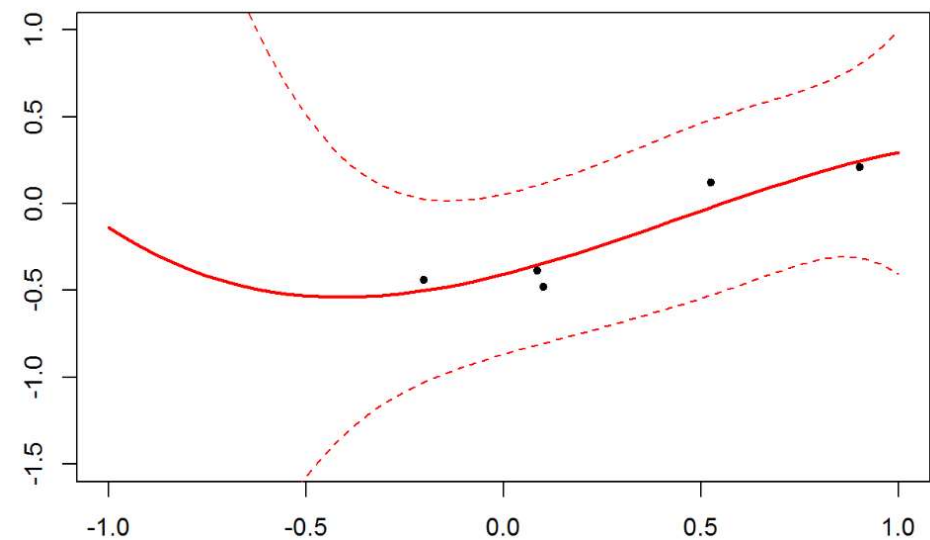
```

#예측모델에 대한 95% 신뢰구간을 return하는 함수
get_predictive_vals <- function(x, m_N, S_N, phi) {
  phix <- sapply(phi, function(base) base(x))
  mean_pred <- t(m_N) %*% phix
  sd_pred <- sqrt(1/beta + t(phix) %*% S_N %*% phix)
  c(mean_pred, mean_pred-2*sd_pred, mean_pred+2*sd_pred)
}

draw_predictive <- function(xs, m_N, S_N, phi) {
  vs <- rep(NA, length(xs))
  ys <- data.frame(means=vs, p2.5=vs, p97.5=vs) # init dataframe
  for (i in 1:length(xs)) { # compute predictive values for all xs
    ys[i,] <- get_predictive_vals(xs[i],m_N, S_N, phi)
  }
  # draw mean and 95% interval
  lines(xs, ys[,1], col="red", lwd=2)
  lines(xs, ys[,2], col="red", lty="dashed")
  lines(xs, ys[,3], col="red", lty="dashed")
}

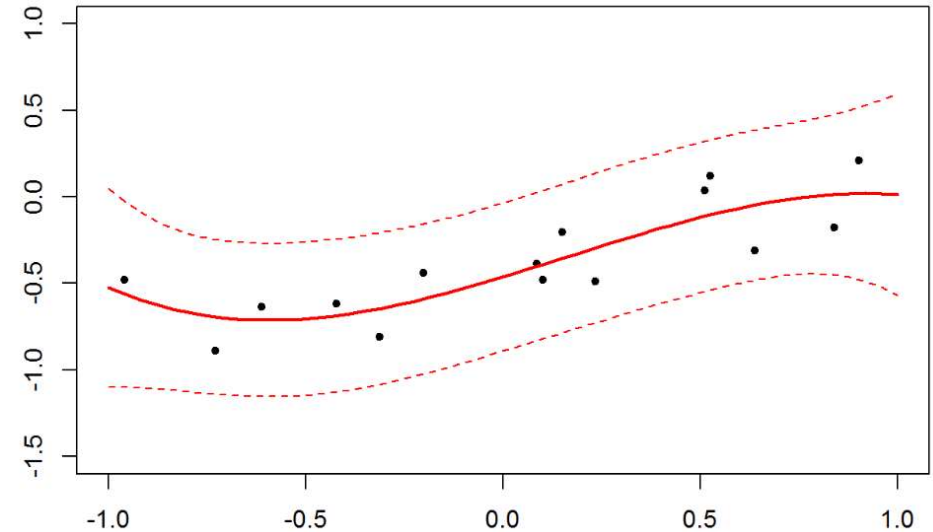
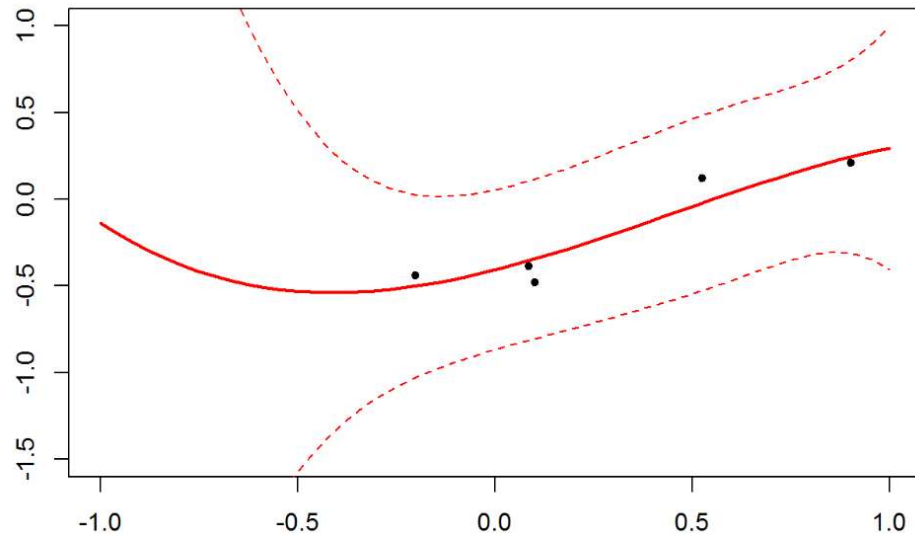
set.seed(121)
X <- make.X(5) # make some points
Y <- make.Y(X)
phi <- c(one,id,sq,x3) # basis for the cubic regression
m_0 <- c(0,0,0,0) # priors
S_0 <- alpha*diag(4)
posterior_1 <- compute_posterior(X, Y, m_0, S_0, phi=phi)
m_N <- posterior_1$m
S_N <- posterior_1$S
plot(X, Y, pch=20, ylim=c(-1.5,1), xlim=c(-1,1), ylab="y", xlab="x")
xs <- seq(-1,1,len=50)
draw_predictive(xs, m_N, S_N, phi=phi)

```



새로운 데이터 추가!

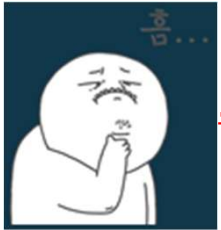
```
X_new <- make.X(10) # more points are available!  
Y_new <- make.Y(X_new)  
posterior_2 <- compute_posterior(X_new, Y_new, posterior_1$m, posterior_1$S, phi=phi)  
m_N <- posterior_2$m  
S_N <- posterior_2$S  
plot(c(X,X_new), c(Y,Y_new), pch=20, ylim=c(-1.5,1), xlim=c(-1,1), ylab="y", xlab="x")  
draw_predictive(xs, m_N, S_N, phi=phi)
```



-1 근처 data들이 들어오면서 예측분포가 더 정확해 졌다.

베이지안 관점에서 모델 비교 문제

- MLE의 과적합 문제는 점추정에서 기여 \rightarrow 주변화를 통해 해결 (합산, 적분 \approx 분포추정)
- 모든 데이터 셋을 Train Set으로 이용 가능하며, Validation Set이 따로 필요하지 않다 (Cross-Validation 불필요)
- 7장. 상관벡터 머신 에서 좀더 자세히 다룰 예정.



모델의 불확실성을 확률로 나타내고 가법 정리 · 곱셈 정리를 이용하여 평가하자

- 사용 하는 변수

새로운 입력 : X

그에 대한 출력 (예측하려는) : t

모델 중 매개 변수 : W

관찰 (교육) 자료 : D

L 개의 모델 : $\{M_i\} \quad (i = 1, 2, \dots, L)$

3.4 Bayesian Model Comparison

모델의 불확실성을 확률로 나타내고 가법 정리 · 곱셈 정리를 이용하여 평가하자

- 모델이 선택될 확률은 $p(M_i)$ 로 표현 되지만 보통 모르기 때문에 $\frac{1}{n}$ 로 둔다

$$p(M_i | D) \propto p(M_i) p(D | M_i)$$

따라서, 베이즈 정리에 의해 전개된 식을 보면 **후자에 더 많은 관심을** 두게 됨.

- $p(D|M_i)$ 는 Model evidence 라고 불리며, 서로 다른 모델들에 대한 데이터로서 보여지는 선호도 주변 가능성도 라고도 불림
- 두 모델에 대한 evidence비율은 $p(D|M_i) / p(D|M_j)$ 는 Bayes factor

3.4 Bayesian Model Comparison - Mixture distribution

모델의 사후 분포 $p(\mathbf{D}|\mathbf{M}_i)$ 를 알면 예측 분포 *predictive distribution*
(새로운 \mathbf{x} 대해 t 가 어떤 값이 될지)도 가법 정리와 곱셈 정리로 표현 가능하다.

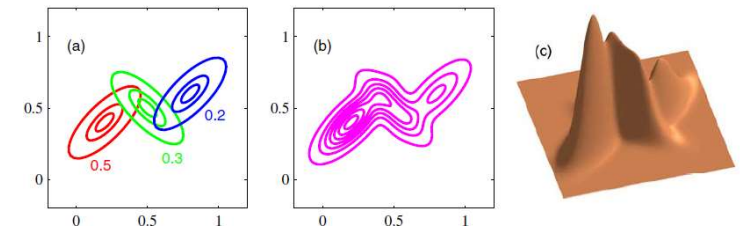
$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^L p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D})$$

겨우 오늘 이해..

Train Set, \mathbf{D} 와 새로운 데이터 \mathbf{X} 가 있을 때, Target \mathbf{t} 는 모델 \mathbf{M} 에 대해 위 식과 같이 전개 됨.

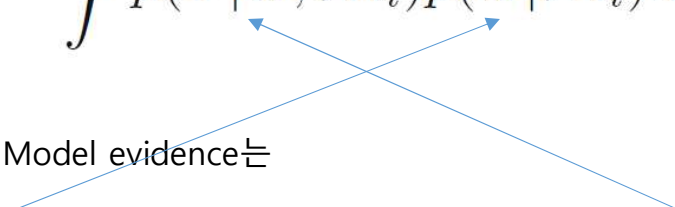
- 왜 이게 혼합 분포일까?

\mathbf{D} 와 \mathbf{x} , 모델 i 로 Target t 를 예측한 다음, 데이터에 알맞은 정도
(Model evidence)를 가중치로 weighted Sum을 해주기 때문



3.4 Bayesian Model Comparison - **Model selection**

파라미터 w 를 가진 모델 M_i 의 evidence를 또한 가법 정리와 곱셈 정리로 분해 해 보면

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}$$


위 식과 같게 되고, 표본추출의 측면에서 Model evidence는

사전 분포로부터 랜덤하게 표본 추출한 w 들을 바탕으로 모델 M 으로 부터 데이터 집합 D 를 생성하게 될 확률로 정의

→ **M 의 parameter w 가 D 를 생성할 확률**

어떤 w 가 선택이 되는지 실험을 해봅시다

3.4 Bayesian Model Comparison - **Model selection**

어떤 하나의 매개 변수 \mathbf{w} 를 가진 모델을 생각한다
 특정 모델 \mathcal{M} (단순회귀-편차가 없음) 을 가정하면

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}$$

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw \simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

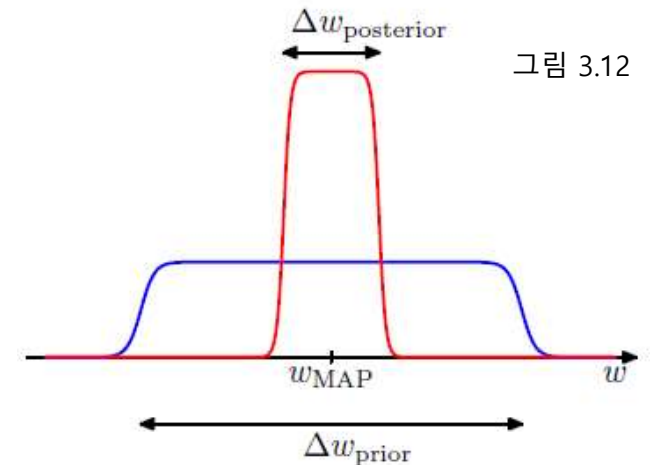


그림 3.12

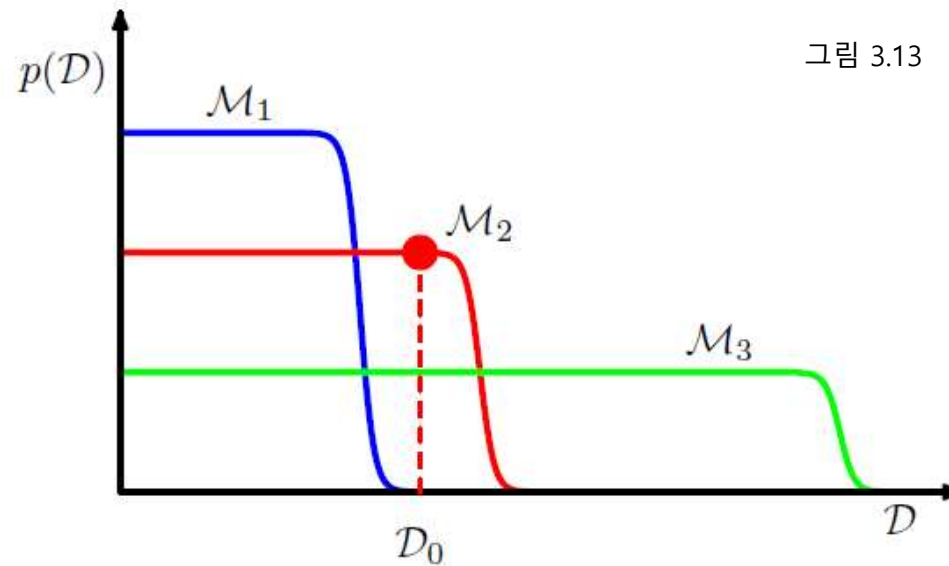
$\int p(\mathcal{D}|w)dw$ 만 본다면, 최고점 w_{MAP} 와 폭 $\Delta w_{\text{posterior}}$ 를 곱하면 적분의 근사값을 구할 수 있다.

Prior가 사각형 모습이고, Δw_{prior} 를 폭으로 가진다면 $p(w)$ 는 위와 같이 됨.

여기에 로그를 취하면 다음과 같이 된다.

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)$$

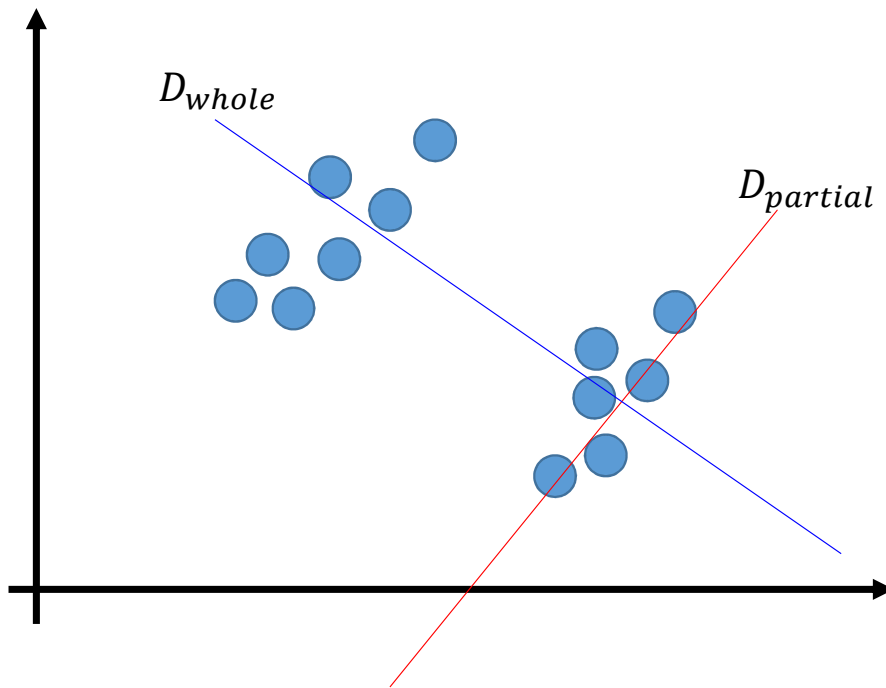
3.4 Bayesian Model Comparison - Model selection



가로축은 데이터 세트가 취할 수 있는 값을 1 차원으로 표현.

- 모델의 복잡성은 $M_1 < M_2 < M_3$ 한다.
1. 간단한 모델 M_1 이 생성 할 수 있는 데이터의 범위가 좁아 (여러 매개 변수를 바꾸어도 비슷한 데이터 세트 밖에 나오지 않는다)
 2. 복잡한 모델 M_3 는 여러가지 데이터를 생성 할 수 있지만 커버하는 D의 범위가 크므로 각각의 발현 확률은 낮다.
 3. 특정 데이터 세트 D_0 에 대해서는 중간 복잡성을 가진 모델 M_2 가 가장 큰 evidence를 가지게 된다.

3.4 Bayesian Model Comparison - Expected Bayes factor



M_1 이 진정한 모델 이라고 한다면, **베이즈 요인**(bayes factor)는 개별 데이터로 보면 잘못된 M_2 로 커지는 경우도 있지만, (Model evidence 가 잘못 적합한 M_2 가 클 수도 있지만)

베이즈 요인은 D 의 분포에서 기댓값을 구하게 되면 우리가 평균적으로 올바른 모델을 구하게 될 것을 보장한다.

$$\int p(\mathcal{D}|\mathcal{M}_1) \ln \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} d\mathcal{D}$$

KL-Divergence와 정확히 일치하며, 항상 0보다 큰 값을 가지므로 **평균적으로 베이즈 요인**은 올바른 모델을 선택한다.

3.4 Bayesian Model Comparison - Summary

- Bayesian framework는 과도한 학습을 피하고 훈련 데이터만 사용해 모델을 비교할 수 있다.
- 하지만 베이esian 역시 모델의 형태에 대한 가정이 필요하고, 그것은 잘못된 결론을 이끌었다
- 결론은 사전 분포의 특성에 상당히 의존
 - 비 사전 분포는 정규화 상수가 정의 할 수 없기 때문에 evidence를 정의 할 수 없다
- 실제 응용에서는 독립적 인 테스트 데이터를 평가 용으로 가지고 푸는 것이 현명 (< 어 결국?)



- <https://heavywatal.github.io/lectures/prml-3-4.html>
- <http://www.di.fc.ul.pt/~jpn/r/PRML/chapter3.html>