

# Tema 5: Modelos Probabilísticos para Patrones Visuales. PPCA

Rafael Molina

A partir del libro de Bishop: PRML

# Contenido

1. Introducción
2. Modelo
3. Estimación de parámetros por máxima verosimilitud
4. Estimación de parámetros usando Inferencia Variacional (VI)

# I. Introducción

La formulación PCA más conocida está basada en la proyección lineal de los datos en un subespacio de menor dimensión.

Sin embargo, existe una formulación muy interesante de las PCA como modelo de variables latentes.

Esta formulación que se llama PCA probabilística (PPCA), tiene varias ventajas sobre la llamada formulación clásica. Dichas ventajas las iremos viendo en este tema.

## II. Modelo

Consideremos la distribución a priori sobre variables latentes  $\mathbf{z}$

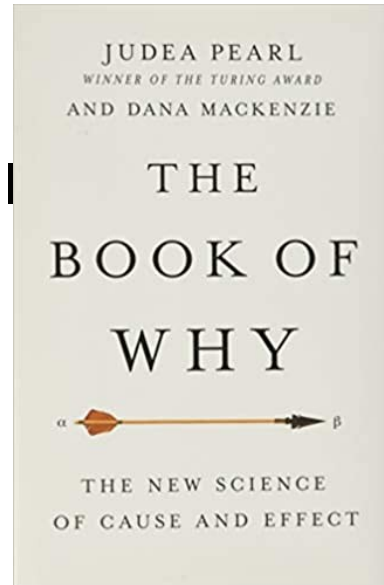
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

y supongamos el modelo de observación

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

$\mathbf{W}$  es de tamaño  $D \times M$  y  $\boldsymbol{\mu}$  es un vector  $D$ -dimensional. El otro parámetro del modelo es un escalar  $\sigma^2$  que gobierna la varianza de la distribución condicionada.

¿Cuál es la interpretación generativa del modelo?



## II. Modelo

¿Serías capaz de explicar la forma alargada de esta distribución?

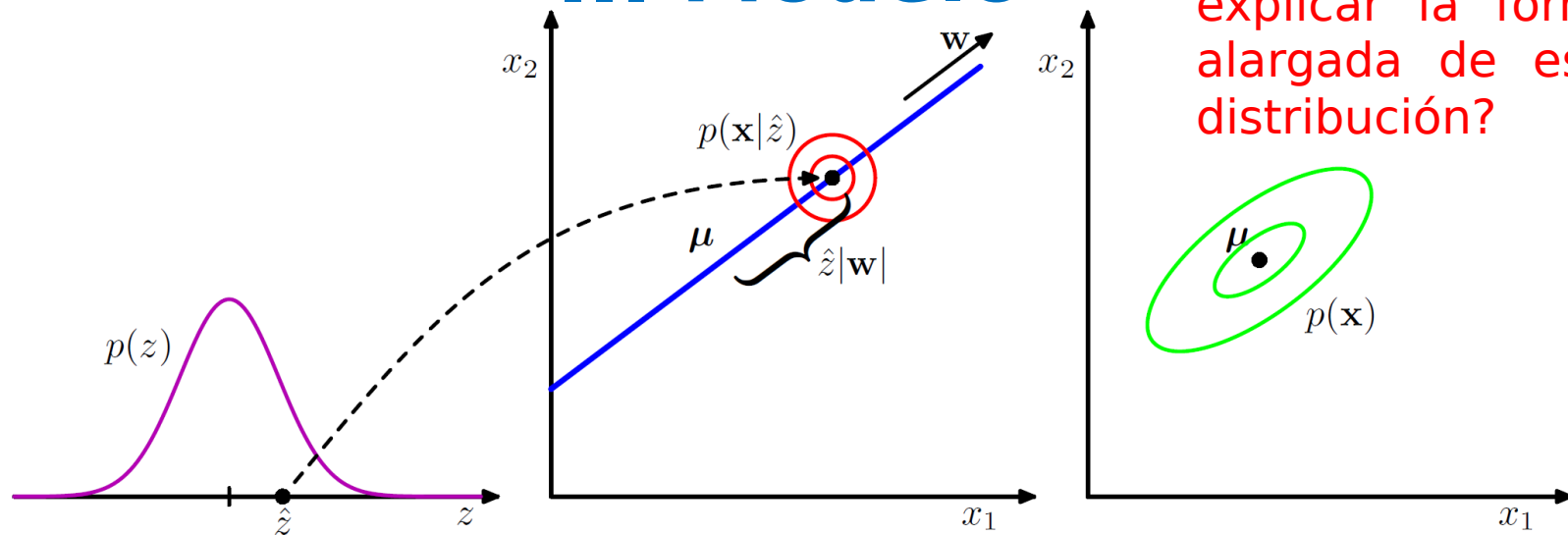


Ilustración de PPCA como modelo generativo para datos bidimensionales y una variable latente. Un dato observado  $\mathbf{x}$  es generado de la forma siguiente:

- primero extraemos un  $\hat{z}$  de la variable latente de la distribución a priori  $p(z)$
- a con  $w\hat{z} + \mu$  extraemos un valor de  $\mathbf{x}$  de una distribución de Gauss isotrópica (círculos en rojo) con media  $w\hat{z} + \mu$  y covarianza  $\sigma^2\mathbf{I}$ .

Las elipses en verde muestran contornos de densidad de la marginal  $p(\mathbf{x})$ .

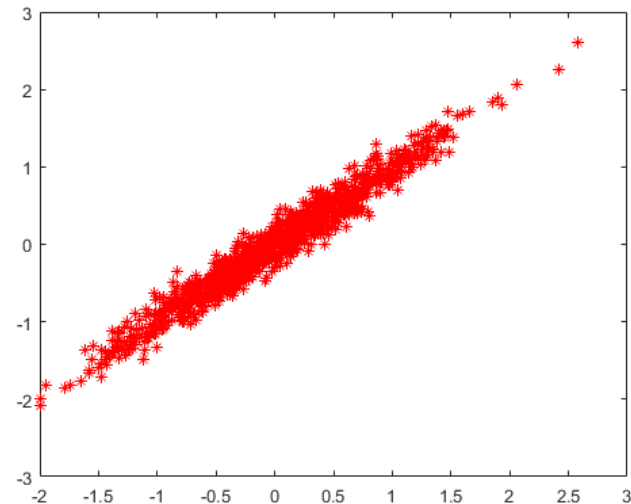
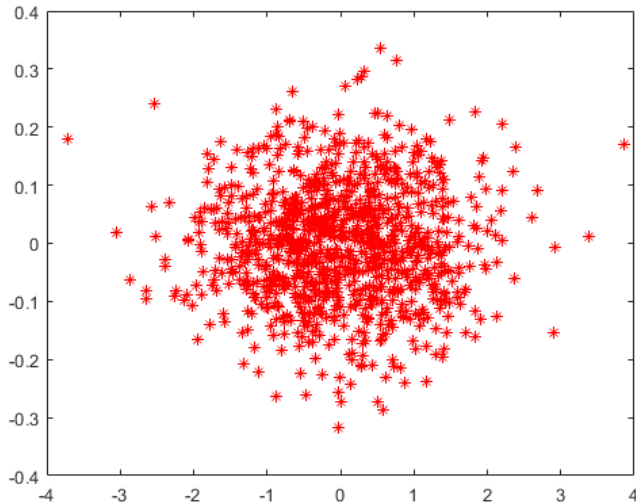
```
n=1000;
z=randn(1,n);
w=[1 0]';
w=w/norm(w);
x=w*z+0.1*randn(2,n);
figure;
plot(x(1,:),x(2,:),'*r');
```

Éste es un modelo generativo típico: generamos un  $z$ , lo multiplicamos por  $W$  y le sumamos un ruido. Obtenemos  $x$ .

```
n=1000;
z=randn(1,n);
w=[1 1]';
w=w/norm(w);
x=w*z+0.1*randn(2,n);
figure;
plot(x(1,:),x(2,:),'*r');
```

Nos permite generar ejemplos de  $z$ , y a partir de ellos generar  $x$ . También nos permite preguntarnos por  $z$  dado un  $x$ .

Para generar ejemplos y *asociar*  $z$  a observaciones necesitamos estimar



# III. Modelo

Para este nuevo modelo de nuestros datos tenemos que estimar los valores de  $\mathbf{W}$ ,  $\mu$  y  $\sigma^2$ . Podemos usar varias aproximaciones:

- Aproximación Moda A Posteriori (MAP): Asignar distribuciones a priori a estos parámetros y encontrar cuales son los de mayor probabilidad a posteriori dadas las observaciones.
- Aproximación Bayesiana: Integrar en las distribuciones de los parámetros.
- Máxima Verosimilitud (ML): Calcular

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) d\mathbf{z}$$

y encontrar los parámetros que hacen más probables nuestras observaciones. **Ésta es la que vamos a seguir.**

# III. Modelo

Podemos probar fácilmente (ya sabemos como hacerlo, mira el tema anterior) que

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C}) \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Observa:

- que las distribuciones condicionada y a priori sean todas normales ayuda mucho en todo el proceso de estimación que vamos a hacer.



# III. Modelo

## Algo a tener en cuenta:

Hay una redundancia en la parametrización correspondiente a rotaciones de las coordenadas del espacio latente. Si consideras la matriz

$$\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$$

donde  $\mathbf{R}$  es una matriz ortonormal y usamos la propiedad  $\mathbf{R}\mathbf{R}^T = \mathbf{I}$  vemos que

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T = \mathbf{W}\mathbf{R}\mathbf{R}^T\mathbf{W}^T = \mathbf{W}\mathbf{W}^T$$

### III. Modelo

Para calcular la **distribución predictiva** (marginal de  $\mathbf{x}$ ) necesitamos calcular la inversa de

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Ésta es una matriz de tamaño  $D \times D$  que puede ser complicado invertir (recuerda que su inversión es de orden  $O(D^3)$ ). Utilizando la identidad de Woodbury de inversión de matrices podemos escribir

$$\mathbf{C}^{-1} = \sigma^{-2} \mathbf{I} - \sigma^{-2} \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^T$$

con

$$\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}$$

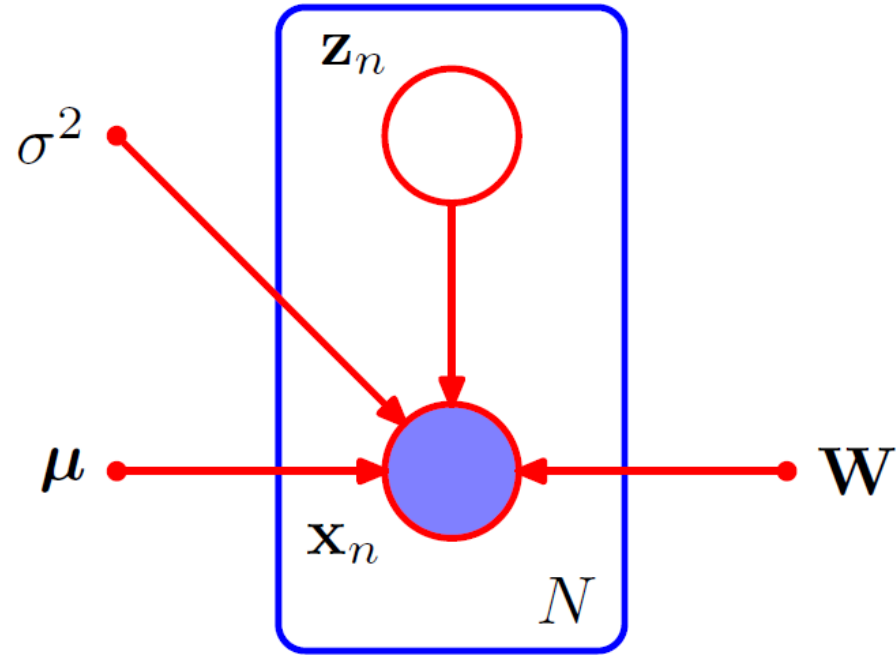
nuestra inversión de matrices se convierte en un problema  $O(M^3)$ .

## IV Estimación de parámetros por máxima verosimilitud

Vamos, por fin, a estimar los parámetros del modelo utilizando el principio de máxima verosimilitud.

Suponemos que tenemos un conjunto  $\mathbf{X} = \{\mathbf{x}_n\}$

de instancias observados (observa que aquí es  $N$  el número de observaciones, en PCAs usamos  $I$  para el número de observaciones). El modelo PPCA



## IV. Estimación de parámetros por máxima verosimilitud

La correspondiente función de verosimilitud es

$$\begin{aligned}\ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \sum_{n=1}^N \ln p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\end{aligned}$$

Si derivamos esta función con respecto a todo lo desconocido, lo primero que puede probarse muy fácilmente es que

$$\boldsymbol{\mu} = \bar{\mathbf{X}}$$

## IV. Estimación de parámetros por máxima verosimilitud

Podemos reescribir el logaritmo de la verosimilitud como

$$\ln p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{N}{2} \left\{ D \ln(2\pi) + \ln |\mathbf{C}| + \text{Tr}(\mathbf{C}^{-1}\mathbf{S}) \right\}$$

donde  $\mathbf{S}$  es la matriz de covarianza muestral definida mediante

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

Obviamente, si realizásemos un promedio sobre muchas muestras tendría

$$\mathbf{E}(\mathbf{S}) = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Lo que nos ayudará  
mucho a entender lo  
que viene

## IV Estimación de parámetros por máxima verosimilitud

La maximización con respecto a  $\mathbf{W}$  y  $\sigma^2$  es más complicada. Tipping and Bishop (1999b) probaron que **todos los puntos estacionarios de la log verosimilitud** pueden escribirse como

$$\mathbf{W}_{\text{ML}} = \mathbf{U}_M (\mathbf{L}_M - \sigma^2 \mathbf{I})^{1/2} \mathbf{R}$$

donde

- $\mathbf{U}_M$  es una matriz  $D \times M$  cuyas columnas son cualquier subconjunto (de tamaño  $M$ ) de los autovectores de la matriz de covarianza muestral  $\mathbf{S}$ ,
- La matriz diagonal  $\mathbf{L}_M$  de tamaño  $M \times M$  contiene los correspondientes autovalores  $\lambda_i$ , de la matriz de covarianza muestral  $\mathbf{S}$  y
- $\mathbf{R}$  es una matriz ortonormal arbitraria de tamaño  $M \times M$

## IV Estimación de parámetros por máxima verosimilitud

Tipping and Bishop (1999b) probaron también que (y esto es lo que nos interesa):

el estimador de máxima verosimilitud de  $\mathbf{W}$  se obtiene cuando los  $M$  autovectores son los asociados a los  $M$  mayores autovalores de la matriz de covarianza muestral (las otras soluciones son puntos de silla).

Supongamos que los autovectores han sido ordenados en orden decreciente de los correspondientes autovalores. Los  $M$  primeros autovectores los vamos a notar  $\mathbf{u}_1, \dots, \mathbf{u}_M$ . En este caso, las columnas de  $\mathbf{W}_{ML}$  definen el llamado subespacio principal standard PCA.

## IV Estimación de parámetros por máxima verosimilitud

Por último.

El correspondiente estimador de máxima verosimilitud (MLE) de la varianza es

$$\sigma_{\text{ML}}^2 = \frac{1}{D - M} \sum_{i=M+1}^D \lambda_i$$



## IV. Estimación de parámetros por máxima verosimilitud

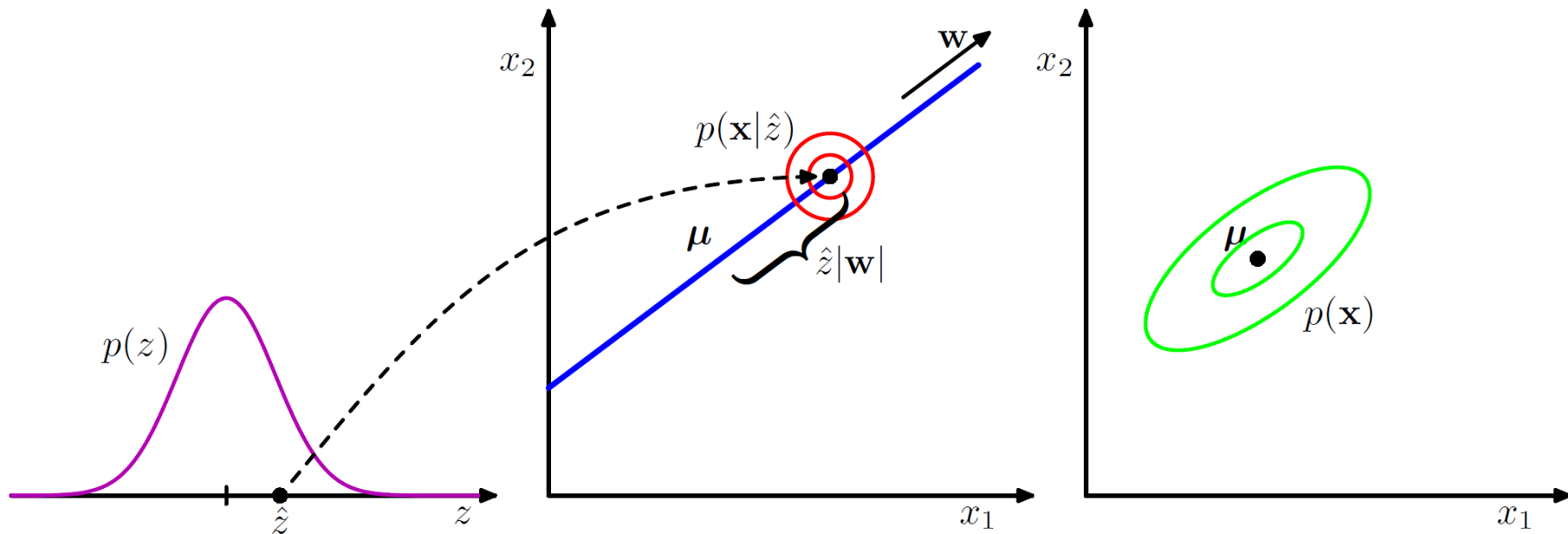
La existencia de una matriz de rotación en la MLE de  $W$  es un hecho importante que no discutiremos aquí. Podéis consultar el libro de Bishop para analizar las implicaciones.

Vamos a dedicar un poco de tiempo a estudiar la matriz de covarianza de la marginal

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

## IV. Estimación de parámetros por máxima verosimilitud

Recuerda el gráfico



## IV. Estimación de parámetros por máxima verosimilitud

Consideremos la varianza de la distribución predictiva en una dirección especificada por el vector  $\mathbf{v}$ , donde  $\mathbf{v}^T \mathbf{v} = 1$ .

Como sabemos, la varianza vale  $\mathbf{v}^T \mathbf{C} \mathbf{v}$ .

- Si  $\mathbf{v}$  es ortogonal al subespacio principal entonces  $\mathbf{v}^T \mathbf{U} = \mathbf{0}$  y, por tanto,  $\mathbf{v}^T \mathbf{C} \mathbf{v} = \sigma^2$ . El modelo predice una varianza que es la media de los autovalores no considerados.
- Si  $\mathbf{v} = \mathbf{u}_i$  donde  $\mathbf{u}_i$  es uno de los autovectores del subespacio principal retenidos entonces  $\mathbf{v}^T \mathbf{C} \mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$ .

En otras palabras, el modelo captura correctamente la varianza de los datos en los ejes principales y la promedia en el resto.

## IV. Estimación de parámetros por máxima verosimilitud

Una vez que hemos calculado los parámetros del modelo podemos calcular la distribución a posteriori  $p(\mathbf{z}|\mathbf{x})$  que tiene la forma

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N} \left( \mathbf{z} | \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2} \mathbf{M} \right)$$

Debería ser -1, compruébal

$$\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}$$

Debería ser +, compruéba

La demostración es muy sencilla. Ver el vídeo último del tema anterior.

## IV. Estimación de parámetros por máxima verosimilitud

PCA convencional se entiende, normalmente, como una proyección de datos mientras que PPCA se formula como una transformación de variables latentes en datos observados.

Para aplicaciones de visualización y compresión de datos, podemos invertir la transformación usando la regla de Bayes y obtenemos para cualquier punto  $\mathbf{x}$

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{\text{ML}}^T(\mathbf{x} - \bar{\mathbf{x}})$$

Observa que  
estoy  
suprimiendo  
ML en M por  
simplicidad

con

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$$

Fíjate que estamos diciendo: si tengo que coger un valor, elijo la media pero ahora tengo otros muchos.

## IV. Estimación de parámetros por máxima verosimilitud

Vamos ahora a estudiar brevemente la relación con PCA.

En el límite ( $\sigma^2 \rightarrow 0$ ), la media a posteriori del modelo PCA probabilístico es

$$(\mathbf{W}_{\text{ML}}^T \mathbf{W}_{\text{ML}})^{-1} \mathbf{W}_{\text{ML}}^T (\mathbf{x} - \bar{\mathbf{x}})$$

que no coincide con el  $\mathbf{z}$  que obtiene PCA standard (en PPCA los ejes están reescalados) pero al multiplicarla por  $\mathbf{W}_{\text{ML}}$  produce la misma proyección. Recuerda que en PCA

no reescalamos

$$\mathbf{W}_{\text{ML}} = \mathbf{U}_M (\mathbf{L}_M - \sigma^2 \mathbf{I})^{1/2} \mathbf{R}$$

## IV. Estimación de parámetros por máxima verosimilitud

En el límite  $\sigma^2 \rightarrow 0$  La covarianza a posteriori en el modelo PCA probabilístico es cero. Recuerda que la covarianza vale  $\sigma^2 \mathbf{M}^{-1}$  con

$$\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}$$

que es la matriz nula cuando  $\sigma^2 = 0$

Observa que para  $\sigma^2 > 0$ , la proyección latente se desplaza hacia el origen en relación a la proyección ortogonal. Recuerda

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1} \mathbf{W}_{\text{ML}}^T (\mathbf{x} - \bar{\mathbf{x}})$$

$$\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}$$

## V. Estimación de parámetros usando Inferencia Variacional (VI)

En esta sección estimaremos iterativamente los parámetros del modelo:  $\boldsymbol{\mu}$ ,  $\mathbf{W}$  y  $\sigma^2$  y al mismo tiempo iremos estimando la distribución a posteriori de  $\mathbf{z}$  dado  $\mathbf{x}$ .

Para PCA probabilístico la estimación de los parámetros se aborda usando el algoritmo EM pero nosotros usaremos inferencia variacional. El algoritmo EM es un caso particular de la inferencia variacional.

Las ideas de esta sección están en la base de muchos modelos de Deep Learning probabilísticos. En particular del que veremos en el tema siguiente, el Autoencoder Variacional (VAE)



# V. Estimación de parámetros usando Inferencia Variacional (VI)

En la estimación MLE habíamos optimizado directamente. Teníamos

$$\begin{aligned}\ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \sum_{n=1}^N \ln p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\end{aligned}$$

y resolvíamos

$$\hat{\mathbf{W}}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2 = \arg \max_{\mathbf{W}, \boldsymbol{\mu}, \sigma^2} \sum_{n=1}^N \ln p(\mathbf{x}_n)$$

Hemos quitado la dependencia de  $p(\mathbf{x}_n)$  de  $\mathbf{W}$ ,  $\boldsymbol{\mu}$  y  $\sigma^2$  por comodidad.

## V. Estimación de parámetros usando Inferencia Variacional (VI)

Sabemos que

$$\ln p(\mathbf{x}_n) = \ln \int p(\mathbf{x}_n, \mathbf{z}_n) d\mathbf{z}_n$$

y que

$$\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^N \{\ln p(\mathbf{x}_n | \mathbf{z}_n) + \ln p(\mathbf{z}_n)\}$$

Desgraciadamente no tenemos los valores de  $\mathbf{z}_n$ . Si los tuviéramos,

## V. Estimación de parámetros usando Inferencia Variacional (VI)

Usamos ahora la desigualdad de Jensen que nos dice que para cualquier distribución  $q(\mathbf{z}_n)$  que queramos utilizar

$$\ln p(\mathbf{x}_n) = \ln \int p(\mathbf{x}_n, \mathbf{z}_n) d\mathbf{z}_n \geq \int q(\mathbf{z}_n) \ln \frac{p(\mathbf{x}_n, \mathbf{z}_n)}{q(\mathbf{z}_n)} d\mathbf{z}_n$$

alcanzando la igualdad solo  $q(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{x}_n)$

## V. Estimación de parámetros usando Inferencia Variacional (VI)

Podemos convertir, entonces, nuestro objetivo en

$$\hat{\mathbf{W}}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2, \hat{p}(\mathbf{z}_1|\mathbf{x}_1), \dots, \hat{p}(\mathbf{z}_N|\mathbf{x}_1) = \arg \max_{\mathbf{W}, \boldsymbol{\mu}, \sigma^2, q(\mathbf{z}_1), \dots, q(\mathbf{z}_N)} \sum_{n=1}^N \int q(\mathbf{z}_n) \ln \frac{p(\mathbf{x}_n, \mathbf{z}_n)}{q(\mathbf{z}_n)} d\mathbf{z}_n$$

Para resolver el problema anterior alternamos entre la optimización en  $\boldsymbol{\mu}$ ,  $\mathbf{W}$  y  $\sigma^2$  y  $q(\mathbf{z}_1), \dots, q(\mathbf{z}_N)$ .

## V. Estimación de parámetros usando Inferencia Variacional (VI)

Dados  $\boldsymbol{\mu}$ ,  $\mathbf{W}$  y  $\sigma^2$ . Ya sabemos que  $\boldsymbol{\mu}$  debe fijarse a la media muestral.

### Paso 1

Puede probarse muy fácilmente que el mejor  $q()$  que podemos elegir para que la anterior integral sea lo más

grā  $q(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n | \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}), \sigma^{-2} \mathbf{M})$  Debe ser  $\sigma^2 \mathbf{M}^{-1}$

Fíjate que esta distribución ya la habíamos obtenido y

# V. Estimación de parámetros usando Inferencia Variacional (VI)

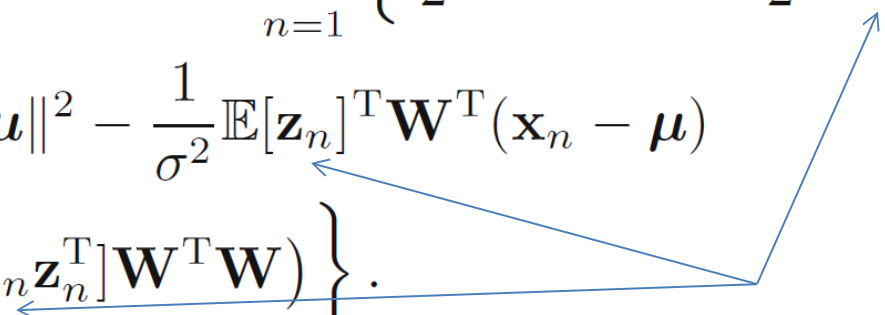
## Paso 2

Ahora fijamos  $q(z_1), \dots, q(z_N)$ , tenemos que maximizar en  $\boldsymbol{\mu}$ ,  $\mathbf{W}$  y  $\sigma^2$ . Observa que  $q()$  esta fija y no hay nada que optimizar en ella). Por tanto calculamos y maximizamos en  $\boldsymbol{\mu}$ ,  $\mathbf{W}$  y  $\sigma^2$

$$\sum_{n=1}^N \int q(\mathbf{z}_n) \ln p(\mathbf{x}_n, \mathbf{z}_n) d\mathbf{z}_n$$

# V. Estimación de parámetros usando Inferencia Variacional (VI)

La integral a maximizar vale

$$\begin{aligned}\mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2)] = & - \sum_{n=1}^N \left\{ \frac{D}{2} \ln(2\pi\sigma^2) + \frac{1}{2} \text{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T]) \right. \\ & + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{z}_n]^T \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}) \\ & \left. + \frac{1}{2\sigma^2} \text{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \mathbf{W}^T \mathbf{W}) \right\}.\end{aligned}$$


Estas esperanzas hay que calcularlas usando  $q(\mathbf{z}_n)$ . No te equivoques, sus parámetros son fijos, ahora no hay que optimizarlos.

## V. Estimación de parámetros usando Inferencia Variacional (VI)

Tenemos

$$\begin{aligned}\mathbb{E}[\mathbf{z}_n] &= \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \\ \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] &= \sigma^2 \mathbf{M}^{-1} + \mathbb{E}[\mathbf{z}_n] \mathbb{E}[\mathbf{z}_n]^T\end{aligned}$$

No te confundas. Aquí los parámetros son fijos. No hay que optimizarlos.

lo que produce la actualización (la media la fi principio)

$$\begin{aligned}\mathbf{W}_{\text{new}} &= \left[ \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}}) \mathbb{E}[\mathbf{z}_n]^T \right] \left[ \sum_{n=1}^N \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \right]^{-1} \\ \sigma_{\text{new}}^2 &= \frac{1}{ND} \sum_{n=1}^N \left\{ \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 - 2 \mathbb{E}[\mathbf{z}_n]^T \mathbf{W}_{\text{new}}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \right. \\ &\quad \left. + \text{Tr} \left( \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] \mathbf{W}_{\text{new}}^T \mathbf{W}_{\text{new}} \right) \right\}.\end{aligned}$$

**Repetimos ahora los pasos 1 y 2 con estos nuevos parámetros.**



## V. Estimación de parámetros usando Inferencia Variacional (VI)

¿Qué extensiones del modelo se te ocurren?

Algunas de ellas las veremos en el capítulo siguiente