Machine Learning for Signal Processing
Linear Gaussian Models

Abelino Jimenez
Logistics

• HW4 Today
• Last quiz This weekend
• Poster session December 4^{th}
  _ Instructions will be posted soon
  _ Instructions for SV and Kigali
• Final Report December 6th
Machine Learning for Signal Processing

Linear Gaussian Models

Abelino Jimenez
Parameters Estimation

A Maximum Likelihood Estimator maximizes

$$\mathbb{P}(\text{data} \mid \text{parameters})$$

A Maximum A Posteriori Estimator maximizes

$$\mathbb{P}(\text{parameters} \mid \text{data})$$

$$\mathbb{P}(\text{parameters} \mid \text{data}) = \frac{\mathbb{P}(\text{data} \mid \text{parameters}) \cdot \mathbb{P}(\text{parameters})}{\mathbb{P}(\text{data})}$$
MAP estimation of continuous variables

An example

\[ x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2) \]

We want to estimate the mean, but we have a previous knowledge

\[ \mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \]

given

\[ \hat{\mu}_{MAP} = \arg \max_{\mu} p(\mu | x_1, \ldots, x_n) \]
MAP estimation of continuous variables

An example

\[ \hat{\mu}_{MAP} = \arg \max_{\mu} p(x_1, \ldots, x_n | \mu) \cdot p(\mu) \]

\[ p(x_1, \ldots, x_n | \mu) \cdot p(\mu) = \prod_{i=1}^{n} p(x_i | \mu) \cdot p(\mu) \]

\[ = \frac{1}{\sqrt{2\pi}\sigma_0} \exp \left( -\frac{1}{2} \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{1}{2} \left( \frac{x_i - \mu}{\sigma_0} \right)^2 \right) \]

MAP minimizes

\[ \sum_{i=1}^{n} \left( \frac{x_i - \mu}{\sigma} \right)^2 + \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \]

\[ \hat{\mu}_{MAP} = \frac{\sigma_0^2 n}{\sigma_0^2 n + \sigma^2} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) + \frac{\sigma^2}{\sigma_0^2 n + \sigma^2} \mu_0 \]

\[ \hat{\mu}_{MAP} \rightarrow \hat{\mu}_{ML} \]
MAP estimation of continuous variables

What is a good prior?

- Analyze the support of the distribution

\[ x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2) \quad \mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \]

\[ \mu | x_1, \ldots, x_n \sim \text{Gaussian} \]

\( p(\mu) \) and \( p(\mu | x_1, \ldots, x_n) \) belong to the same class of distributions.
MAP estimation of continuous variables

\[ x_1, \ldots, x_n \sim \text{Bernoulli}(p) \]

We can consider

\[ p \sim \text{Beta}(\alpha, \beta) \]

\[ p \mid x_1, \ldots, x_n \sim \text{Beta} \left( \alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i \right) \]

\[ \hat{p}_{MAP} = \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{\alpha + \beta + n - 2} \]

We say that the Beta distribution is the conjugate prior of the Bernoulli distribution.
# Conjugate Prior

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Model parameters</th>
<th>Conjugate prior distribution</th>
<th>Prior hyperparameters</th>
<th>Posterior hyperparameters</th>
<th>Interpretation of hyperparameters[^note 1]</th>
<th>Posterior predictive[^note 2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bemoulli</td>
<td>$p$ (probability)</td>
<td>Beta</td>
<td>$\alpha, \beta$</td>
<td>$\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i$</td>
<td>$\alpha - 1$ successes, $\beta - 1$ failures[^note 1]</td>
<td>$p(\hat{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$p$ (probability)</td>
<td>Beta</td>
<td>$\alpha, \beta$</td>
<td>$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$</td>
<td>$\alpha - 1$ successes, $\beta - 1$ failures[^note 1]</td>
<td>BetaBin($\hat{x}$</td>
</tr>
<tr>
<td>Negative binomial</td>
<td>$p$ (probability)</td>
<td>Beta</td>
<td>$\alpha, \beta$</td>
<td>$\alpha + \sum_{i=1}^{n} x_i, \beta + rn$</td>
<td>$\alpha - 1$ total successes, $\beta - 1$ failures[^note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming $r$ stays fixed)</td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td>$\lambda$ (rate)</td>
<td>Gamma</td>
<td>$k, \theta$</td>
<td>$k + \sum_{i=1}^{n} x_i, \frac{\theta}{n\theta + 1}$</td>
<td>$k$ total occurrences in $\frac{1}{\theta}$ intervals</td>
<td>NB($\hat{x}$</td>
</tr>
<tr>
<td>Categorical</td>
<td>$p$ (probability vector), $k$ (number of categories; i.e., size of $p$)</td>
<td>Dirichlet</td>
<td>$\alpha$</td>
<td>$\alpha + (c_1, \ldots, c_k)$, where $c_i$ is the number of observations in category $i$</td>
<td>$\alpha_i - 1$ occurrences of category $i$[^note 1]</td>
<td></td>
</tr>
<tr>
<td>Multinomial</td>
<td>$p$ (probability vector), $k$ (number of categories; i.e., size of $p$)</td>
<td>Dirichlet</td>
<td>$\alpha$</td>
<td>$\alpha + \sum_{i=1}^{n} x_i$</td>
<td>$\alpha_i - 1$ occurrences of category $i$[^note 1]</td>
<td>DirMult($\hat{x}$</td>
</tr>
<tr>
<td>Hypergeometric</td>
<td>$M$ (number of target members)</td>
<td>Beta-binomial[^4]</td>
<td>$n = N, \alpha, \beta$</td>
<td>$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$</td>
<td>$\alpha - 1$ successes, $\beta - 1$ failures[^note 1]</td>
<td></td>
</tr>
<tr>
<td>Geometric</td>
<td>$p_0$ (probability)</td>
<td>Beta</td>
<td>$\alpha, \beta$</td>
<td>$\alpha + n, \beta + \sum_{i=1}^{n} x_i - n$</td>
<td>$\alpha - 1$ experiments, $\beta - 1$ total failures[^note 1]</td>
<td></td>
</tr>
</tbody>
</table>

[^note 1]: For more details on the interpretation of hyperparameters, refer to the relevant sources.

[^note 2]: The posterior predictive distribution is derived from the conjugate prior and the likelihood function.
Probabilistic Linear Regression

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon = \beta^T X + \varepsilon, \]

where

\[ \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad X = (1 \ X_1 \ X_2 \ \cdots \ X_p)^T \quad \text{and} \quad \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \]

\[
p(y \mid \beta) = p(y_1, y_2, \ldots, y_N \mid \beta) = \prod_{i=1}^{N} p(y_i \mid \beta) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid \beta^T x_i, \sigma^2)
\]

\[
\ell(\beta) = \log \left( \prod_{i=1}^{N} \mathcal{N}(y_i \mid \beta^T x_i, \sigma^2) \right) = \sum_{i=1}^{N} \log \left( \mathcal{N}(y_i \mid \beta^T x_i, \sigma^2) \right)
= \sum_{i=1}^{N} \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right) \right)
= N \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{i=1}^{N} \frac{(y_i - \beta^T x_i)^2}{2\sigma^2}.
\]

\[
\hat{\beta}_{\text{ML}} = \arg \max_{\beta} \ell(\beta) = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2,
\]

\[ \beta \text{ is an unknown parameter not random} \]
A problem with regressions

- ML fit is sensitive
  - Error is squared
  - Small variations in data $\rightarrow$ large variations in weights
  - Outliers affect it adversely
- Unstable
  - If dimension of $X$ $\geq$ no. of instances
    - $(XX^T)^{-1}$ is not invertible

$$A = (XX^T)^{-1} XY^T$$
Map estimation of weights

- Assume weights drawn from a Gaussian
  \[ P(a) = \mathcal{N}(0, \sigma^2 I) \]
- Max. Likelihood estimate
  \[ \hat{a} = \arg \max_a \log P(Y \mid X; a) \]
- Maximum a posteriori estimate
  \[ \hat{a} = \arg \max_a \log P(a \mid Y, X) = \arg \max_a \log P(Y \mid X, a) P(a) \]
MAP estimate priors

- Left: Gaussian Prior on $W$
- Right: Laplacian Prior
MAP estimate of weights

\[ \hat{\beta}_{\text{MAP}} = \arg \max_{\beta} p(\beta \mid y) = \arg \max_{\beta} p(y \mid \beta)p(\beta) = \arg \max_{\beta} (\log p(y \mid \beta) + \log p(\beta)) \]

\[ p(\beta) = \mathcal{N}(\beta \mid 0, \alpha I_p). \]

\[ \hat{\beta}_{\text{MAP}} = \arg \max_{\beta} \prod_{i=1}^{N} \mathcal{N}(y_i \mid \beta^T x_i, \sigma^2) \mathcal{N}(\beta \mid 0, \alpha I_p) \]

\[ = \arg \max_{\beta} - \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 - \alpha \sum_{i=1}^{p} \beta_i^2 \]

\[ = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \sigma^2 \alpha \| \beta \|^2. \]

Loss + Regularization
MAP estimate of weights

\[ \hat{\beta}_{\text{MAP}} = \left( \alpha I_p + \sigma^{-2} X^T X \right)^{-1} \sigma^{-2} X^T y = \left( X^T X + \sigma^2 \alpha I_p \right)^{-1} X^T y, \]

- Equivalent to *diagonal loading* of correlation matrix
  - Improves condition number of correlation matrix
    - Can be inverted with greater stability
  - Will not affect the estimation from well-conditioned data
  - Also called Tikhonov Regularization
    - Dual form: Ridge regression

- **MAP estimate of weights**
  - Not to be confused with MAP estimate of Y
MAP estimate of weights

$$\hat{\beta}_{MAP} = \arg\max_\beta p(\beta | y) = \arg\max_\beta p(y | \beta)p(\beta) = \arg\max_\beta (\log p(y | \beta) + \log p(\beta))$$

$$\beta \sim \text{Laplace}(0, bI)$$

$$\hat{\beta}_{MAP} = \arg\min_\beta \sum_{i=1}^{N} (y_i - x_i^\top \beta)^2 + \sigma^2 \sigma \|\beta\|_1$$

Loss + Regularization
MAP estimation of weights with Laplacian prior

• Assume weights drawn from a Laplacian
  \[ P(a) = \lambda^{-1} \exp(-\lambda^{-1}|a|_1) \]
  – Maximum a posteriori estimate

\[ \hat{a} = \arg \max_a \ C' - (y - a^T X)^T (y - a^T X)^T - \lambda^{-1}|a|_1 \]

• No closed form solution
  – Quadratic programming solution required
    • Non-trivial
MAP estimation of weights with Laplacian prior

- Assume weights drawn from a Laplacian
  \[ P(a) = \lambda^{-1}\exp(-\lambda^{-1}|a|_1) \]
- Maximum a posteriori estimate
  \[ \hat{a} = \arg\max_a C'(y - a^T X)^T (y - a^T X)^T - \lambda^{-1}|a|_1 \]
- Identical to $L_1$ regularized least-squares estimation
GAUSSIAN CASE
Joint Gaussian

• \( x \) and \( y \) are jointly Gaussian

\[
\begin{align*}
    z &= \begin{bmatrix} x \\ y \end{bmatrix} \\
    E[z] &= \mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \\
    C_{xy} &= E[(x - \mu_x)(y - \mu_y)^T] \\
    \text{Var}(z) &= C_{zz} = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} \\
    P(z) &= N(\mu_z, C_{zz}) = \frac{1}{\sqrt{2\pi} \left| C_{zz} \right|} \exp\left( -0.5(z - \mu_z)^T C_{zz}^{-1} (z - \mu_z) \right)
\end{align*}
\]

• \( z \) is Gaussian
MAP estimation: Gaussian PDF

$P(X, Y)$
MAP estimation: The Gaussian at a particular value of $X$
Conditional Probability of $y \mid x$

$$P(y \mid x) = N(\mu_y + C_{yx}C_{xx}^{-1}(x - \mu_x), C_{yy} - C_{yx}C_{xx}^{-1}C_{xy})$$

$$E_{y\mid x}[y] = \mu_{y\mid x} = \mu_y + C_{yx}C_{xx}^{-1}(x - \mu_x)$$

$$Var(y \mid x) = C_{yy} - C_{yx}C_{xx}^{-1}C_{xy}$$

- The conditional probability of $y$ given $x$ is also Gaussian
  - The slice in the figure is Gaussian
- The mean of this Gaussian is a function of $x$
- The variance of $y$ reduces if $x$ is known
  - Uncertainty is reduced
MAP estimation: The Gaussian at a particular value of X

Most likely value
MAP Estimation of a Gaussian RV

\[ \hat{y} = \arg \max_y P(y \mid x) = E_{y \mid x}[y] \]
The conditional probability of $y$ given $x$ is also Gaussian:

$$P(y | x) = N(\mu_y + C_{yx} C_{xx}^{-1} (x - \mu_x), C_{yy} - C_{yx} C_{xx}^{-1} C_{xy})$$

- The slice in the figure is Gaussian.
- The mean of this Gaussian is a function of $x$.
- The variance of $y$ reduces if $x$ is known.
  - Uncertainty is reduced.

**MAP Estimate.**

**Its actually a regression line.**
Conditional Probability of $y | x$

\[
P(y | x) = N(\mu_y + C_{yx} C_{xx}^{-1} (x - \mu_x), C_{yy} - C_{yx} C_{xx}^{-1} C_{xy})
\]

\[
E_{y|x}[y] = \mu_{y|x} = \mu_y + C_{yx} C_{xx}^{-1} (x - \mu_x)
\]

\[
Var(y|x) = C_{yy} - C_{yx} C_{xx}^{-1} C_{xy}
\]

- The conditional probability of $y$ given $x$ is also Gaussian.
  - The slice in the figure is Gaussian.
  - The mean of this Gaussian is a function of $x$.
  - The variance of $y$ reduces if $x$ is known.
    - Uncertainty is reduced.

\[
\mu, \mu_x, \mu_y, \mu_{y|x}, \mu_{y|y} = \mu
\]

\[
\sigma^2, \sigma_x^2, \sigma_y^2, \sigma_{y|x}^2 = 1
\]

The variance of $Y$ shrinks because we know $X$.

Note that the actual value of $X$ doesn't matter. Simply knowing $X$ reduces the variance of $Y$ if the two are correlated.
MMSE estimation

Mean value
Its also a **minimum-mean-squared error** estimate

- Minimize error:

\[
Err = E[\|y - \hat{y}\|^2 | x] = E[(y - \hat{y})^T (y - \hat{y}) | x]
\]

\[
Err = E[y^T y + \hat{y}^T \hat{y} - 2\hat{y}^T y | x] = E[y^T y | x] + \hat{y}^T \hat{y} - 2\hat{y}^T E[y | x]
\]

- Differentiating and equating to 0:

\[
d.Err = 2\hat{y}^T d\hat{y} - 2E[y | x]^T d\hat{y} = 0
\]

\[
\hat{y} = E[y | x]
\]

The MMSE estimate is the mean of the distribution
For the Gaussian: $\text{MAP} = \text{MMSE}$

Most likely value is also the MEAN value.

- Would be true of any symmetric distribution
LINEAR GAUSSIAN MODEL
Remember Eigenfaces

- Approximate every face $f$ as
  $$ f = w_{f,1} V_1 + w_{f,2} V_2 + w_{f,3} V_3 + \ldots + w_{f,k} V_k $$
- Estimate $V$ to minimize the squared error

- **Error is unexplained by $V_1 \ldots V_k$**
- **Error is orthogonal to Eigenfaces**
Eigen Representation

- K-dimensional representation
  - Error is orthogonal to representation
  - Weight and error are specific to data instance

Illustration assuming 3D space
• K-dimensional representation
  – Error is orthogonal to representation
  – Weight and error are specific to data instance

\[ w_1 + e_2 = W_{12} \]
• K-dimensional representation
  – Error is orthogonal to representation

All data with the same representation $wV_1$ lie a plane orthogonal to $wV_1$
With 2 bases

\[ w_{11} + w_{21} + \varepsilon_1 \]

- Error is at 90° to the eigenfaces

- Illustration assuming 3D space

- K-dimensional representation
  - Error is orthogonal to representation
  - Weight and error are specific to data instance
With 2 bases

\[ \mathbf{w}_1 + \mathbf{w}_2 + \mathbf{e}_2 \]

Illustration assuming 3D space

- K-dimensional representation
  - Error is orthogonal to representation
  - Weight and error are specific to data instance

Error is at 90° to the eigenfaces
In Vector Form

\[ X_i = w_{1i} V_1 + w_{2i} V_2 + \varepsilon_i \]

- K-dimensional representation
  - Error is orthogonal to representation
  - Weight and error are specific to data instance

Error is at 90° to the eigenfaces
In Vector Form

\[ X_i = w_{1i} V_1 + w_{2i} V_2 + \varepsilon_i \]

\[ x = Vw + e \]

- \( K \)-dimensional representation
- \( x \) is a \( D \) dimensional vector
- \( V \) is a \( D \times K \) matrix
- \( w \) is a \( K \) dimensional vector
- \( e \) is a \( D \) dimensional vector

Error is at 90° to the eigenface
Constraints

\[ x = Vw + e \]

- \( V^T V = I \) : Eigen vectors are orthogonal to each other
- For every vector, error is orthogonal to Eigen vectors
  - \( e^T V = 0 \)
- Over the collection of data
  - Average \( w w^T = \text{Diagonal} \) : Eigen representations are uncorrelated
  - \( e^T e = \text{minimum} \) : Error variance is minimum
    - Mean of error is 0

Error is at 90° to the eigenface
A Statistical Formulation of PCA

\[ x = Vw + e \]

\[ w \sim N(0, B) \]

\[ e \sim N(0, E) \]

- \( x \) is a random variable generated according to a linear relation
- \( w \) is drawn from an \( K \)-dimensional Gaussian with diagonal covariance
- \( e \) is drawn from a 0-mean \((D-K)\)-rank \( D \)-dimensional Gaussian
- Estimate \( V \) (and \( B \)) given examples of \( x \)
Linear Gaussian Models!!

\[ x = Vw + e \]
\[ w \sim N(0, B) \]
\[ e \sim N(0, E) \]

- \( x \) is a random variable generated according to a linear relation
- \( w \) is drawn from a Gaussian
- \( e \) is drawn from a 0-mean Gaussian
- Estimate \( V \) given examples of \( x \)
  - In the process also estimate \( B \) and \( E \)
Linear Gaussian Models!!

- \( x \) is a random variable generated according to a linear relation
- \( w \) is drawn from a Gaussian
- \( e \) is drawn from a zero-mean Gaussian

Estimate \( V \) given examples of \( x \)

- In the process also estimate \( B \) and \( E \)

PCA is a specific instance of a linear Gaussian model with particular constraints
- \( B = \text{Diagonal} \)
- \( \nu^T v = 1 \)
- \( E \) is low rank
Linear Gaussian Models

\[ \mathbf{x} = \mu + \mathbf{Vw} + \mathbf{e} \quad \mathbf{w} \sim N(0, \mathbf{B}) \quad \mathbf{e} \sim N(0, \mathbf{E}) \]

• Observations are linear functions of two *uncorrelated* Gaussian random variables
  – A “weight” variable \( \mathbf{w} \)
  – An “error” variable \( \mathbf{e} \)
  – Error not correlated to weight: \( \mathbb{E}[\mathbf{e}^T \mathbf{w}] = 0 \)

• Learning LGMs: Estimate parameters of the model given instances of \( \mathbf{x} \)
  – The problem of learning the distribution of a Gaussian RV
LGMs: Probability Density

\[ x = \mu + Vw + e \]
\[ w \sim N(0, B) \]
\[ e \sim N(0, E) \]

• The mean of \( x \):

\[ E[x] = \mu + V E[w] + E[e] = \mu \]

• The Covariance of \( x \):

\[ E[(x - E[x])(x - E[x])^T] = VBV^T + E \]
The probability of $x$

$x = \mu + Vw + e$

$w \sim N(0, B)$

$e \sim N(0, E)$

$x \sim N(\mu, VBV^T + E)$

$P(x) = \frac{1}{\sqrt{(2\pi)^D | VBV^T + E |}} \exp\left(-0.5(x - \mu)^T \left(VBV^T + E\right)^{-1}(x - \mu)\right)$

- $x$ is a linear function of Gaussians: $x$ is also Gaussian
- Its mean and variance are as given
Estimating the variables of the model

\[ x = \mu + Vw + e \]

\[ w \sim N(0, B) \]
\[ e \sim N(0, E) \]

\[ x \sim N(\mu, VBVT + E) \]

- Estimating the variables of the LGM is equivalent to estimating \( P(x) \)
  - The variables are \( \mu, V, B \) and \( E \)
Estimating the model

\[ x = \mu + Vw + e \]

\[ w \sim N(0, B) \]

\[ e \sim N(0, E) \]

\[ x \sim N(\mu, VBV^T + E) \]

• The model is indeterminate:
  – \( Vw = VCC^{-1}w = (VC)(C^{-1}w) \)
  – We need extra constraints to make the solution unique

• Usual constraint: \( B = I \)
  – Variance of \( w \) is an identity matrix
Estimating the variables of the LGM is equivalent to estimating $P(x)$.

- The variables are $\mu$, $V$, and $E$. 

\[
x = \mu + Vw + e
\]

\[
x \sim N(\mu, VV^T + E)
\]

\[
w \sim N(0, I)
\]

\[
e \sim N(0, E)
\]
The Maximum Likelihood Estimate

\[ x \sim N(\mu, VV^T + E) \]

- Given training set \( x_1, x_2, \ldots, x_N \), find \( \mu, V, E \)

- The ML estimate of \( \mu \) does not depend on the covariance of the Gaussian

\[ \mu = \frac{1}{N} \sum_{i} x_i \]
Centered Data

• We can safely assume “centered” data
  – $\mu = 0$

• If the data are not centered, “center” it
  – Estimate mean of data
    • Which is the maximum likelihood estimate
  – Subtract it from the data
Simplified Model

\[ x = Vw + e \]

\[ w \sim N(0, I) \]

\[ e \sim N(0, E) \]

\[ x \sim N(0, VV^T + E) \]

- Estimating the variables of the LGM is equivalent to estimating \( P(x) \)
  - The variables are \( V \), and \( E \)
Estimating the model

\[ x = Vw + e \]

\[ x \sim N(0, VV^T + E) \]

- Given a collection of \( x_i \) terms
  - \( x_1, x_2, \ldots, x_N \)
- Estimate \( V \) and \( E \)
- \( w \) is unknown for each \( x \)
- But if assume we know \( w \) for each \( x \), then what do we get:
Estimating the Parameters

\[ x_i = Vw_i + e \]

\[ P(e) = N(0, E) \]

\[ P(x \mid w) = N(Vw, E) \]
Reminder: $x$ and $w$ are jointly Gaussian

$$x = Vw + e$$

$$P(x) = N(0, VV^T + E)$$

$$P(w) = N(0, I)$$

$$C_{xw} = E[(x - \mu_x)(w - \mu_w)^T] = V$$

$$z = \begin{bmatrix} x \\ w \end{bmatrix}$$

$$P(z) = N(\mu_z, C_{zz})$$

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_w \end{bmatrix} = 0$$

$$C_{zz} = \begin{bmatrix} C_{xx} & C_{xw} \\ C_{wx} & C_{ww} \end{bmatrix}$$

$$C_{zz} = \begin{bmatrix} VV^T + E & V \\ V^T & I \end{bmatrix}$$

• $x$ and $w$ are jointly Gaussian!
MAP estimation: Gaussian PDF
MAP estimation: The Gaussian at a particular value of $X$
Conditional Probability of $x|w$

$$P(x|w) = N(\mu_x + C_{xw}C_{ww}^{-1}(w - \mu_w), C_{xx} - C_{xw}C_{ww}^{-1}C_{wx})$$

$$= N(C_{xw}C_{ww}^{-1}w, C_{xx} - C_{xw}C_{ww}^{-1}C_{wx})$$

$$E_{x|w}[x] = C_{xw}C_{ww}^{-1}w$$

$$\text{Var}(x|w) = C_{xx} - C_{xw}C_{ww}^{-1}C_{wx}$$

• Comparing to

$$P(x|w) = N(Vw, E)$$

• We get:

$$V = C_{xw}C_{ww}^{-1}$$

$$E = C_{xx} - C_{xw}C_{ww}^{-1}C_{wx}$$
Or more explicitly

\[ C_{ww} = \frac{1}{N} \sum_i w_i w_i^T \]

\[ C_{xw} = \frac{1}{N} \sum_i x_i w_i^T \]

\[ V = C_{xw} C_{ww}^{-1} \]

\[ E = C_{xx} - C_{xw} C_{ww}^{-1} C_{wx} \]

\[ V = \left( \sum_i x_i w_i^T \right) \left( \sum_i w_i w_i^T \right)^{-1} \]

\[ E = \frac{1}{N} \left( \sum_i x_i x_i^T - V \sum_i w_i x_i^T \right) \]
Estimating LGMs: If we know $w$

$$x_i = Vw_i + e$$

$P(e) = N(0, E)$

$$V = \left( \sum_i x_iw_i^T \right) \left( \sum_i w_iw_i^T \right)^{-1}$$

$$E = \frac{1}{N} \left( \sum_i x_ix_i^T - V \sum_i w_ix_i^T \right)$$

- But in reality we *don’t* know the $w$ for each $x$
  – So how to deal with this?

- EM..
Recall EM

• We figured out how to compute parameters if we *knew* the missing information
• Then we “fragmented” the observations according to the posterior probability $P(z|x)$ and counted as usual
• In effect we took the expectation with respect to the a posteriori probability of the missing data: $P(z|x)$
EM for LGMs

\[ x_i = Vw_i + e \]

\[ P(e) = N(0, E) \]

\[ V = \left( \sum_i x_i w_i^T \right) \left( \sum_i w_i w_i^T \right)^{-1} \]

\[ E = \frac{1}{N} \left( \sum_i x_i x_i^T - V \sum_i w_i x_i^T \right) \]

- Replace unseen data terms with expectations taken w.r.t. \( P(w|x_i) \)
**EM for LGMs**

\[ x_i = Vw_i + e \]

\[ P(e) = N(0, E) \]

\[
V = \left( \sum_i x_i w_i^T \right) \left( \sum_i w_i w_i^T \right)^{-1}
\]

\[
E = \frac{1}{N} \left( \sum_i x_i x_i^T - V \sum_i w_i w_i^T \right)
\]

- Replace unseen data terms with expectations taken w.r.t. \( P(w|x_i) \)
Flipping the problem

• How do we estimate the above terms?
• MAP to the rescue!!

\[ E_{w|x_i}[w] \]

\[ E_{w|x_i}[ww^T] \]
Expected Value of $w$ given $x$

$x = Vw + e$

$P(e) = N(0, E)$

$P(w) = N(0, I)$

$P(x) = N(0, VV^T + E)$

• $x$ and $w$ are jointly Gaussian!
  – $x$ is Gaussian
  – $w$ is Gaussian
  – They are linearly related

$z = \begin{bmatrix} x \\ w \end{bmatrix}$

$P(z) = N(\mu_z, C_{zz})$
Recall: $w$ and $x$ are jointly Gaussian.

$$x = Vw + e$$

$e \sim N(0, E)$ \hspace{1cm} $P(w) = N(0, I)$

$P(x) = N(0, VV^T + E)$

$C_{xx} = VV^T + E$ \hspace{1cm} $C_{ww} = I$

$C_{xw} = E[(x - \mu_x)(w - \mu_w)^T] = V$

$z = \begin{bmatrix} x \\ w \end{bmatrix}$

$P(z) = N(\mu_z, C_{zz})$

$\mu_z = \begin{bmatrix} \mu_x \\ \mu_w \end{bmatrix} = 0$

$C_{zz} = \begin{bmatrix} C_{xx} & C_{xw} \\ C_{wx} & C_{ww} \end{bmatrix}$

• $x$ and $w$ are jointly Gaussian!
Recall: \( w \) and \( x \) are jointly Gaussian

\[
C_{zz} = \begin{bmatrix}
(VV^T + E) & V \\
V^T & I
\end{bmatrix}
\]

\[
C_{xx} = VV^T + E \quad C_{ww} = I
\]

\[
C_{xw} = E[(x - \mu_x)(w - \mu_w)^T] = V
\]

\[
z = \begin{bmatrix}
x \\
w
\end{bmatrix}
\]

\[
P(z) = N(\mu_z, C_{zz})
\]

\[
\mu_z = \begin{bmatrix}
\mu_x \\
\mu_w
\end{bmatrix} = 0
\]

\[
C_{zz} = \begin{bmatrix}
C_{xx} & C_{xw} \\
C_{wx} & C_{ww}
\end{bmatrix}
\]

• \( x \) and \( w \) are jointly Gaussian!
\( P(w \mid z) \)

- \( P(w \mid z) \) is a Gaussian

\[
P(w \mid x) = N(\mu_w + C_{wx} C_{xx}^{-1} (x - \mu_x), C_{ww} - C_{wx} C_{xx}^{-1} C_{xw})
\]

\[
= N(C_{wx} C_{xx}^{-1} x, C_{ww} - C_{wx} C_{xx}^{-1} C_{xw})
\]

\[
= N(V^T (VV^T + E)^{-1} x, I - V^T (VV^T + E)^{-1} V)
\]

\[
Var(w \mid x) = I - V^T (VV^T + E)^{-1} V
\]

\[
E_{w \mid x_i}[w] = V^T (VV^T + E)^{-1} x_i
\]

\[
E_{w \mid x_i}[ww^T] = Var(w \mid x) + E_{w \mid x_i}[w]E_{w \mid x_i}[w]^T
\]

\[
E_{w \mid x_i}[ww^T] = I - V^T (VV^T + E)^{-1} V + E_{w \mid x_i}[w]E_{w \mid x_i}[w]^T
\]
LGM: The complete EM algorithm

\[ \mathbf{x} = \mathbf{Vw} + \mathbf{e} \quad \mathbf{e} \sim \mathcal{N}(0, \mathbf{E}) \quad P(\mathbf{w}) = \mathcal{N}(0, \mathbf{I}) \]

\[ P(\mathbf{x}) = \mathcal{N}(0, \mathbf{VV}^T + \mathbf{E}) \]

- Initialize \( \mathbf{V} \) and \( \mathbf{E} \)
- E step:
  \[ E_{w|x_i}[\mathbf{w}] = \mathbf{V}^T (\mathbf{VV}^T + \mathbf{E})^{-1} \mathbf{x}_i \]
  \[ E_{w|x_i}[\mathbf{ww}^T] = \mathbf{I} - \mathbf{V}^T (\mathbf{VV}^T + \mathbf{E})^{-1} \mathbf{V} + E_{w|x_i}[\mathbf{w}] E_{w|x_i}[\mathbf{w}]^T \]
- M step:
  \[ \mathbf{V} = \left( \sum_i \mathbf{x}_i E_{w|x_i}[\mathbf{w}^T] \right) \left( \sum_i E_{w|x_i}[\mathbf{ww}^T] \right)^{-1} \]
  \[ E = \frac{1}{N} \sum_i \mathbf{x}_i \mathbf{x}_i^T - \frac{1}{N} \mathbf{V} \sum_i E_{w|x_i}[\mathbf{w}] \mathbf{x}_i^T \]
So what have we achieved

• Employed a complicated EM algorithm to learn a Gaussian PDF for a variable $x$

• What have we gained???
  – PCA
    • Sensible PCA
    • EM algorithms for PCA (Probabilistic PCA)

• Next class:
  – Factor Analysis
    • FA for feature extraction


LGMs: Application 1

Learning principal components

\[ \mathbf{x} = \mathbf{Vw} + \mathbf{e} \]
\[ \mathbf{w} \sim \mathcal{N}(0, I) \]
\[ \mathbf{e} \sim \mathcal{N}(0, \mathbf{E}) \]

- Find directions that capture most of the variation in the data
- Error is orthogonal to these variations
The full covariance matrix of a Gaussian has $D^2$ terms.

- Fully captures the relationships between variables.
- Problem: Needs a lot of data to estimate robustly.

LGMs: Application 2

Learning with insufficient data