Convex Optimization Theory and Applications

Topic 7 - Parameter Estimation

Li Li

Department of Automation
Tsinghua University

Fall, 2009-2021.

7.0. Outline

7.1. MLE Parameter Estimation
- 7.1.1 Maximum Likelihood Estimation
- 7.1.2 Linear Measurements with iid Noise
- 7.1.3 Nonlinear Measurements with iid Noise

7.2. EM Parameter Estimation
- 7.2.1 GMM Parameter Estimation
- 7.2.2 The Theory of EM

7.3. Kullback-Leibler Divergence

7.4. PPCA

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

"What! You have solved it already?

Well, that would be too much to say. I have discovered a suggestive fact, that is all."

"Winwood Reade is good upon the subject. He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

- *The Sign of Four*, 1890, by Sir Arthur Conan Doyle

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

Suppose we have a density function $f(x|\theta)$ that is governed by the parameter set θ (usually, we had known the exact form of f). The likelihood function L is the function obtained by reversing the roles of x and θ. That means we view θ as the variable and x as the given information

$$L(\theta|x) = f(x|\theta) \tag{7.1}$$

Definition 2.1 In the method of maximum likelihood, we will try to find a value $g(x)$ of the parameter θ that maximizes $L(\theta|x)$ for each x in X. Here $g(x)$ is called a maximum likelihood estimator of θ.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

$L(\theta \mid x)$ can be any function attains its maximum of θ. Given the observed values x_1, …, x_n, the likelihood of θ is

$$L(\theta \mid x) = f(x_1,...,x_n \mid \theta) \tag{7.2}$$

Let's assume that these data are independent and identically distributed (i.i.d.). Thus, the resulting density for the samples can be written as

$$L(\theta \mid x) \cong \prod_{i=1}^{n} f(x_i \mid \theta) \tag{7.3}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

Based on the definition, we will try to find the values of the parameters that have most likely produced the data observed, which leads to

$$\max_{\theta \in \Theta} L(\theta \mid x) \tag{7.4}$$

Notice that the natural logarithm function $\ln(\cdot)$ is strictly increasing. If the maximum value of $L(\theta \mid x)$ does exist, it will occur at the same points as that of $\ln\left[L(\theta \mid x)\right]$. This function is called the log likelihood function and in many cases is easier to work out than the likelihood function, since the density $f(x \mid \theta)$ usually has a product structure.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For one parameter case, if $L(\theta \mid x)$ is differentiable, we get the solution $\hat{\theta}$ by letting (why?)

$$\frac{\partial L(\theta \mid x)}{\partial \theta} = \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i, \theta) = 0 \tag{7.5}$$

and check whether

$$\left. \frac{\partial^2 L(\theta \mid x)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} < 0 \tag{7.6}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For k parameters case, if $L(\theta \mid x)$ is differentiable, we get the solution $\hat{\theta}$ by letting

$$\frac{\partial \ln L(\theta \mid x)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \ln f(x_i, \theta) = 0, \quad j = 1,2,\cdots,k = 0 \qquad (7.7)$$

and check whether the associating Hessian matrix $U = \left\{ U_{ij}(\hat{\theta}) \right\}$ is negative definite

$$U_{ij}(\hat{\theta}) = \left. \frac{\partial^2 \ln L(\theta \mid x)}{\partial \theta_i \partial \theta_j} \right|_{\theta = \hat{\theta}} \qquad (7.8)$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For example, for the observed Poisson process x_i with

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{7.9}$$

the log likelihood is

$$L(\theta) = \ln\left[\prod_{i=1}^{n} f(x_i \mid \theta)\right] = \sum_{i=1}^{n}\left[x_i \ln \lambda - \lambda - \ln(x_i!)\right]$$

$$= \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i) \tag{7.10}$$

7.1. MLE Parameter Estimation

7.1.2 Maximum Likelihood Estimation

We can see that natural logarithm is used against exponentinal function. Study the derivative of $L(\theta)$, we find the maximum as

$$\frac{dL}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \tag{7.11}$$

which implies the estimator below, since $\left.\frac{\partial^2 l}{\partial \lambda}\right|_{\lambda=\hat{\lambda}} = -\frac{1}{\hat{\lambda}^2}\sum_{i=1}^{n} x_i < 0$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7.12}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For the observed process x_i with the normal distribution as

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7.13}$$

the log likelihood should be

$$L(\theta) = \ln\left[\prod_{i=1}^{n} f(x_i \mid \theta)\right] = -n\ln\sigma - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{7.14}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

The partial derivative yields and thus the MLE estimators are

$$\begin{cases} \dfrac{\partial L}{\partial \mu} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \dfrac{\partial L}{\partial \sigma} = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases} \tag{7.15}$$

$$\begin{cases} \hat{\mu} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma} = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2} \end{cases} \tag{7.16}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

At this point, we have

$$\left[\begin{array}{cc} \frac{\partial^2 L}{\partial^2 \mu} & \frac{\partial^2 L}{\partial \mu \partial \sigma} \\ \frac{\partial^2 L}{\partial \sigma \partial \mu} & \frac{\partial^2 L}{\partial^2 \sigma} \end{array}\right]\Bigg|_{\begin{cases} \hat{\mu}=\frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2} \end{cases}} = \left[\begin{array}{cc} -\frac{n^2}{\sum_{i=1}^{n}(x_i-\hat{\mu})^2} & 0 \\ 0 & -\frac{2n^2}{\sum_{i=1}^{n}(x_i-\hat{\mu})^2} \end{array}\right] < 0 \qquad (7.17)$$

So the estimator above is correct.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

Suppose we have m cells/boxes labeled as 1 up to m. The probability that a ball falls into a box is different (think of the boxes being bigger or smaller). Drop n balls, we record the number of the balls in these cells/boxes as a vector $(x_1,...,x_m)$.

Clearly, the probability that the number of balls in each box is NOT fully independent, the joint probability of a vector $(x_1,...,x_m)$ is called the multinomial. If we have only two cells/boxes, we get binomial distribution.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For the observed multinomial cell probabilities x_i with

$$f(x_1,...,x_m \mid p_1,...,p_m) = \binom{n}{x_1,...,x_m} p_1^{x_1} ... p_m^{x_m} \tag{7.18}$$

Introducing the Lagrange multiplier with coefficient λ, the log likelihood should be

$$L(\theta) = l(\theta) + g(\theta) = \ln n! - \sum_{i=1}^{m} \ln x_i ! + \sum_{i=1}^{m} x_i \ln p_i + \lambda(1 - \sum_{i=1}^{m} p_i) \tag{7.19}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

Without constraint, it is an unsolvable problem in closed form. Notice the parameter space is the simplex

$$\Theta = \left\{ p_i : p_i \geq 0, \sum_{i=1}^{m} p_i = 1 \right\} \subset R^m \tag{7.20}$$

the extremes of $L(\theta)$ should satisfies

$$\nabla L(\theta) = \left(\frac{x_1}{p_1}, \ldots, \frac{x_n}{p_n} \right) = \lambda \nabla g(\theta) = \lambda \tag{7.21}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

Notice $\sum_{i=1}^{m} p_i = 1$, we have

$$\hat{p}_i = \frac{x_i}{n} \tag{7.22}$$

Different from the above example, it is a constrained convex programming problem.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For Gamma distribution

$$f(x_1,...,x_n \mid \alpha,\lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} \tag{7.23}$$

the log likelihood is

$$L(\theta) = \sum_{i=1}^{n} \left[\alpha \ln \lambda + (\alpha - 1) \ln x_i - \lambda x_i - \ln \Gamma(\alpha) \right] \tag{7.24}$$

Is this a convex function?

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

We can obtain that

$$\lambda = \frac{n\alpha}{\sum_{i=1}^{n} x_i} \tag{7.25}$$

$$\ln(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{7.26}$$

What should we do?

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

We can obtain that

$$\lambda = \frac{n\alpha}{\sum_{i=1}^{n} x_i} \tag{7.25}$$

$$\ln(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{7.26}$$

There is no closed-form solution for α and thus for λ. Only numerical solution is available.

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For uniform distribution

$$f(\theta \mid x) = \theta^{-1}, \quad 0 \le x \le \theta \tag{7.27}$$

7.1. MLE Parameter Estimation

7.1.1 Maximum Likelihood Estimation

For uniform distribution

$$f(x \mid \theta) = \theta^{-1}, \quad 0 \le x \le \theta \tag{7.27}$$

the log likelihood is

$$L(x \mid \theta) = \theta^{-n}, \quad 0 \le x_{(1)} \le \ldots \le x_{(n)} \le \theta \tag{7.28}$$

Sketching L will show that it is maximized at $\hat{\theta} = x_{(n)}$. Notice that since this likelihood function has a discontinuity, the maximum is attained at a boundary point.

7.1. MLE Parameter Estimation

7.1.2 Linear Measurements with iid Noise

$$y_i = a_i^T x + v_i, \quad i = 1,\ldots,m \tag{7.29}$$

where $x \in R^n$ is an unknown parameter vector. $a_i \in R^n$, v_i is iid measurement noise with p.d.f. $p(z)$, y_i is the measurement $y \in R^n$ has density $p_x(y) = \prod_{i=1}^{m} p(y_i - a_i^T x)$. Thus, the maximum likelihood estimate is any solution x of

$$L(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x) \tag{7.30}$$

7.1. MLE Parameter Estimation

7.1.2 Linear Measurements with iid Noise

Gaussian noise $N(0,\sigma^2): p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma)^2}$,

$$L(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(a_i{}^T x - y_i)^2 \qquad (7.31)$$

ML estimate is LS solution.

Laplacian noise: $p(z) = (1/(2a))e^{-|z|/a}$,

$$L(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^{m}\left|a_i{}^T x - y_i\right| \qquad (7.32)$$

7.1. MLE Parameter Estimation

7.1.2 Linear Measurements with iid Noise

ML estimate is l_1-norm solution.

Uniform noise on $[-a, a]$, $p(z) = 1/(2a)$

$$L(x) = \begin{cases} -m\log(2a) & \left|a_i^T x - y_i\right| \le a, i = 1, \ldots, m \\ -\infty & \text{otherwise} \end{cases} \qquad (7.33)$$

ML estimate is any x with $\left|a_i^T x - y_i\right| \le a$.

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

Given a Gaussian Mixture Model

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \tag{7.34}$$

The log-likelihood function is

$$L(x) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k) \right\} \tag{7.35}$$

We cannot directly calculate it. Why?

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

Suppose we have n sampled datapoint. Initially, we had a guess of parameters (π_k, μ_k, Σ_k) as $(\pi_k^m, \mu_k^m, \Sigma_k^m)$.

The E-step: estimate the probability of x_i that is generated by the k th Gaussian component

$$\gamma^m(i,k) = \frac{\pi_k^m N(x_i \mid \mu_k^m, \Sigma_k^m)}{\sum_{k=1}^{K} \pi_k^m N(x_i \mid \mu_k^m, \Sigma_k^m)} \tag{7.36}$$

$\gamma^m(i,k)x_i$ can then be viewed as a whole point (the i th point) that is generated by the k th Gaussian component.

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

The M-step: totally, we have n sampled datapoints of $\gamma^m(i,k)x_i$ that are generated by the k th Gaussian component. We can use MLE to get a new guess of parameters (π_k, μ_k, Σ_k)

$$\mu_k^{m+1} = \frac{1}{N_k}\sum_{i=1}^{n}\gamma^m(i,k)x_i, \quad \Sigma_k^{m+1} = \frac{1}{N_k}\sum_{i=1}^{n}\gamma^m(i,k)(x_i-\mu_k^m)(x_i-\mu_k^m)^T \tag{7.37}$$

$$N_k = \sum_{i=1}^{n}\gamma^m(i,k), \quad \pi_k^{m+1} = N_k / n \tag{7.38}$$

Do iterations until it converges.

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

Let use further consider how GMM works.

Step1. Choose a Gaussian component according to the probability π_k and then

Step2. Produce a sample point $x_i \in R^n$ follow the k th Gaussian distribution.

Induce a hidden variable $z_i \in R^n$

$$z_i(k) = \begin{cases} 1, & \text{if } x_i \text{ is produced by } k\text{th Gaussian component} \\ 0, & \text{otherwise} \end{cases} \tag{7.39}$$

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

Then, the likelihood function will be built on $p(x,z)$

$$p(x,z) = p(z)p(x \mid z) \tag{7.40}$$

where $p(x \mid z) = \prod_{k=1}^{K} N(x \mid \mu_k, \Sigma_k)^{z(k)}$ is the PDF of a single Gaussian component and $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ is the probability of the hidden variable.

Now, the new log-likelihood function is

$$L(x,z) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_i(k) \left\{ \log \pi_k + \log N(x_i \mid \mu_k, \Sigma_k) \right\} \tag{7.41}$$

7.2. EM Parameter Estimation

7.2.1 GMM Parameter Estimation

The new log-likelihood function (7.52) can be appropriately optimized, except for one thing: the hidden variable z_i cannot be observed. How do we know about the hidden varible z_i ?

Here, we use the expectation of z_i instead.

$$E\{z_i(k)\} = p(z_i(k)=1 \mid x_i) = \frac{p(z_i(k)=1)\,p(x_i \mid z_i(k)=1)}{p(x_i)}$$

$$= \frac{\pi_k N(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i \mid \mu_j, \Sigma_j)} = \gamma(i,k) \qquad (7.42)$$

7.2. EM Parameter Estimation

7.2.2 The Theory of EM

Suppose x_i represents the observed sample point, and z_i means the corresponding hidden variable. Usually we cannot directly get the parameter θ in likelihood function $p(x|\theta)$ via MLE, but easily optimize the full likelihood function $p(x,z|\theta)$ - that is what EM do.

Induce the distribution of the hidden vaiable $q(z)$. According to Bayesian rule, we have

$$p(x|\theta) = \sum_z q(z) \frac{p(x,z|\theta)}{p(z|x,\theta)} \tag{7.43}$$

7.2. EM Parameter Estimation

7.2.2 The Theory of EM

So, the full log likelihood function can be rewriten as

$$\log p(x \mid \theta) = \log \sum_z q(z) \cdot \frac{p(x, z \mid \theta)}{q(z)} \cdot \frac{q(z)}{p(z \mid x, \theta)} \tag{7.44}$$

Using Jensen's inequality, we have $E[f(x)] \leq f(E[x])$, thus

$$\begin{aligned} \log p(x \mid \theta) &\geq \sum_z q(z) \log \frac{p(x, z \mid \theta)}{q(z)} + \sum_z \left(-q(z) \log \frac{p(z \mid x, \theta)}{q(z)} \right) \\ &= L(q, \theta) + KL(q \,||\, p) \end{aligned} \tag{7.45}$$

7.2. EM Parameter Estimation

7.2.2 The Theory of EM

The term $KL(q \| p)$ is called Kullback-Leibler divergence, describing the difference between distribution p and q. The KL divergence is always greater than or equal to zero, thus

$$L(q,\theta) \leq \log p(x \mid \theta) \tag{7.46}$$

During E-step, we choose $q(z)$ to maximize likelihood function $L(q,\theta)$. The equality of (7.46) holds only when KL divergence is zero, which means the distribution $p(z \mid x,\theta)$ and $q(z)$ are the same. Now, we have

$$q(z) = p(z \mid x,\theta^{(m)}) \tag{7.47}$$

7.2. EM Parameter Estimation

7.2.2 The Theory of EM

During the following M-step, we fix $q(z)$ and choose $\theta^{(m+1)}$ to maximize $L(q,\theta)$. Now $L(q,\theta)$ will be

$$L(q,\theta) = \sum_z p(z \mid x, \theta^{(m)}) \log p(x, z \mid \theta^{(m+1)})$$

$$- \sum_z p(z \mid x, \theta^{(m)}) \log p(z \mid x, \theta^{(m)}) \qquad (7.48)$$

Clearly, the second part in (7.48) keeps constant during M-step, since it is independent with θ. So, we only need to pay attention to the equivalance between the first term of (7.48).

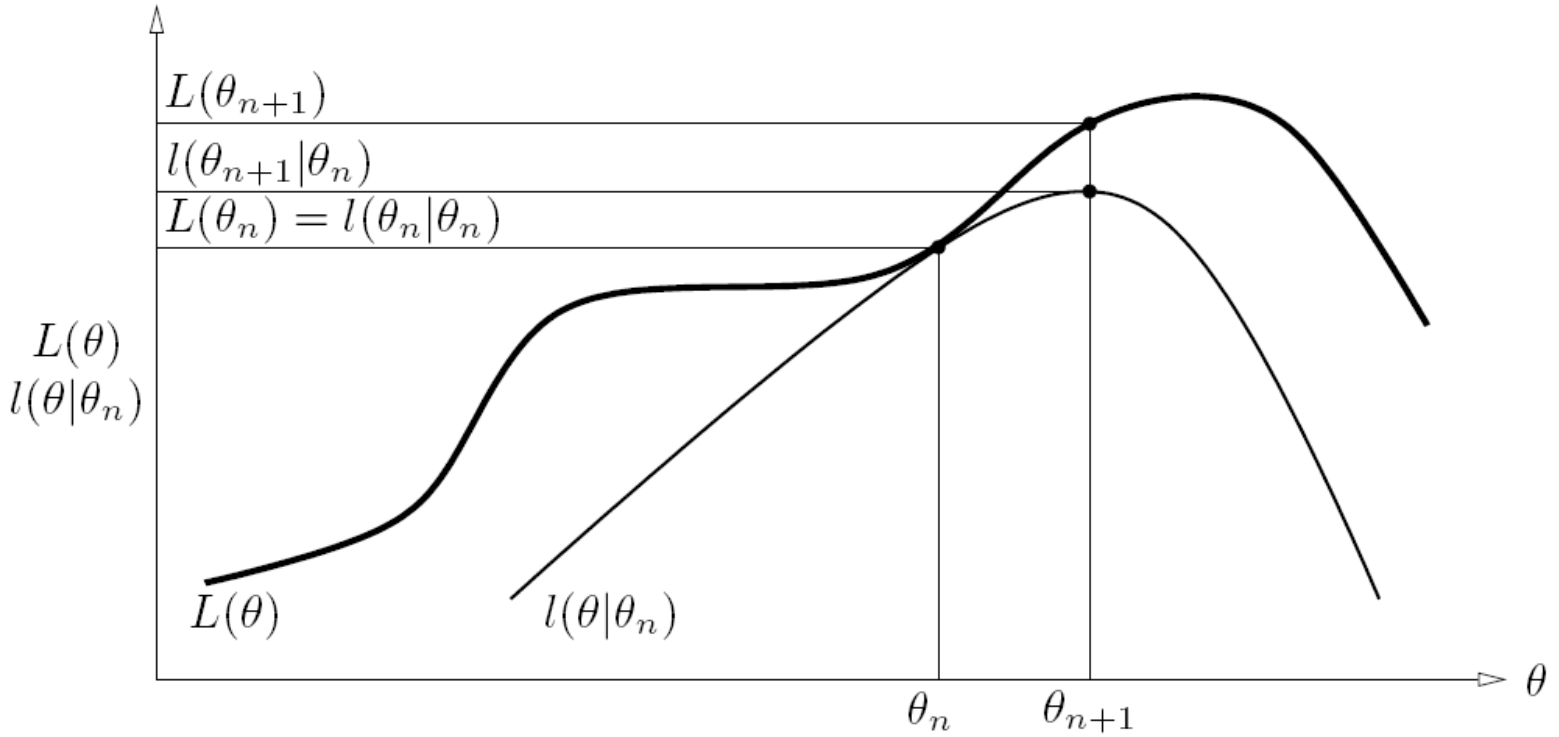
7.2. EM Parameter Estimation

7.2.2 The Theory of EM

We use MLE or any other alternative method to optimize the first part in (7.48).

Because of the above mentioned monotonicity property, the convergence of EM algorithm may be proved under mild regularity conditions.

7.2. EM Parameter Estimation



The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

7.3. Kullback-Leibler Divergence

Gibbs' inequality: suppose that $P = \{p_1, \ldots, p_n\}$ is a probability distribution. For another probability distribution $Q = \{q_1, \ldots, q_n\}$, the following inequality between positive quantities (since the p_i and q_i are positive numbers less than one) holds

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i \qquad (7.49)$$

with equality if and only if $p_i = q_i$ for all i.

That is, the information entropy of a distribution P is less than or equal to its cross entropy with any other distribution Q.

7.3. Kullback-Leibler Divergence

Proof: Note that the natural logarithm satisfies

$$\ln x \le x - 1 \tag{7.50}$$

for all x with equality if and only if $x = 1$.

Let I denote the set of all i for which p_i is non-zero. Then, we have

$$-\sum_{i \in I} p_i \log \frac{q_i}{p_i} \ge -\sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1 \right) = -\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} p_i = -\sum_{i=1}^{n} q_i + 1 \ge 0 \tag{7.51}$$

7.3. Kullback-Leibler Divergence

So, $-\sum_{i\in I} p_i \ln q_i \geq -\sum_{i\in I} p_i \ln p_i$. Thus, $-\sum_{i=1}^{n} p_i \ln q_i \geq -\sum_{i=1}^{n} p_i \ln p_i$, since both hand sides do not grow.

For equality to hold, we require $\frac{q_i}{p_i} = 1$ for all $i \in I$.

The result can be alternatively proved via Jensen's inequality. How to do that?

7.3. Kullback-Leibler Divergence

The difference between the two quantities is the negative of the Kullback-Leibler divergence or relative entropy, so the inequality can also be written:

$$KL(P \| Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0 \qquad (7.52)$$

Note that the use of base-2 logarithms is optional, and allows one to refer to the quantity on each side of the inequality as an "average surprisal" measured in bits.

The Kullback-Leibler divergence is always non-negative as a result known as Gibbs' inequality.

7.3. Kullback-Leibler Divergence

For distributions P and Q of a continuous random variable, KL-divergence is defined to be the integral

$$KL(P \| Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx \geq 0 \qquad (7.53)$$

where $p(x)$ and $q(x)$ denote the densities of P and Q.

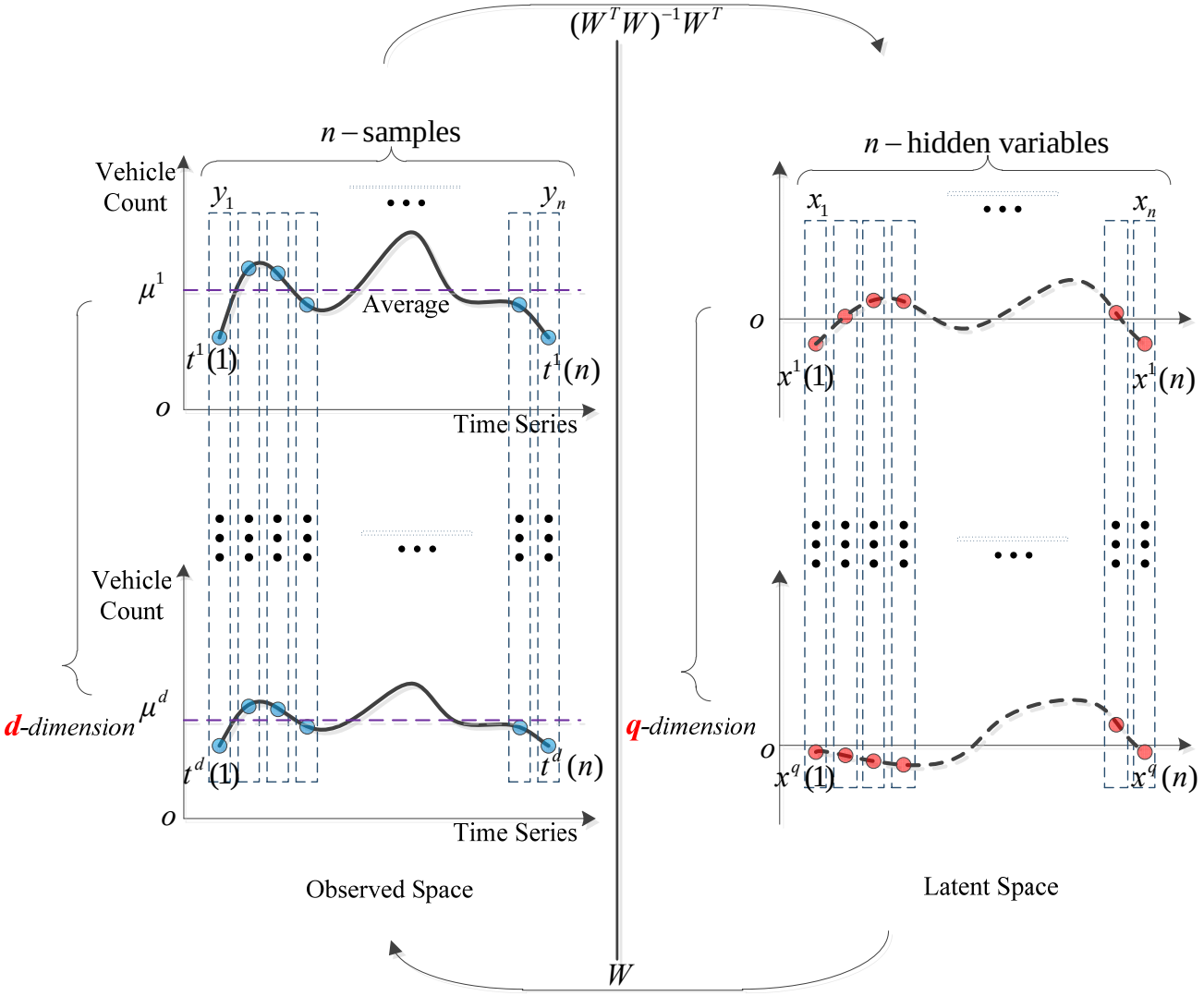
7.4. PPCA

Probabilistic PCA (PPCA) is an extension of original PCA. It supposes that the observed data (traffic flow time series) depend on some latent Gaussian variables as follows

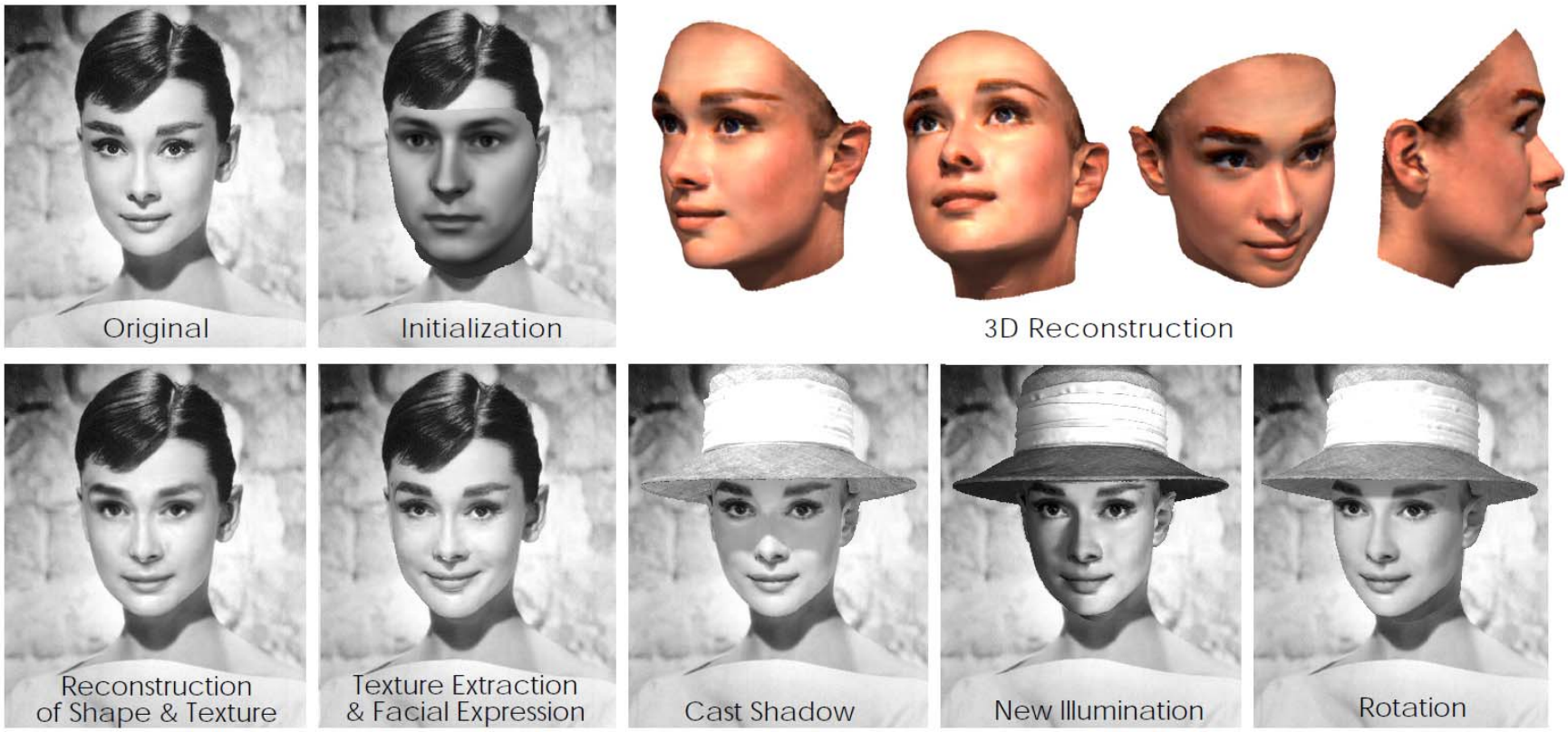
$$t = Wx + \mu + \varepsilon \tag{7.54}$$

where t is a d-dimensional vector of observed data, x is the q-dimensional vector of latent variables ($q < d$), $x \sim N(0, I)$. W is the $d \times q$ projection matrix. The parameter μ permits the model to have a non-zero mean shift; and ε is the isotropic noise $\varepsilon \sim N(0, \sigma^2 I)$.

7.4. PPCA


$(W^TW)^{-1}W^T$
$n-$samples
$n-$hidden variables
Vehicle Count
y_1
y_n
μ^1
Average
$t^1(1)$
$t^1(n)$
o
Time Series
x_1
x_n
$x^1(1)$
$x^1(n)$
Vehicle Count
d-dimension
μ^d
$t^d(1)$
$t^d(n)$
Time Series
q-dimension
$x^q(1)$
$x^q(n)$
Observed Space
Latent Space
W

7.4. PPCA



The figure is copied from "A Morphable Model for the Synthesis of 3D Faces", http://dl.acm.org/citation.cfm?id=311556

7.4. PPCA

The conditional pdf of t given x is given by

$$p(t \mid x, W, \sigma^2, \mu) = N(t \mid Wx + \mu, \sigma^2 I) \tag{7.55}$$

Thus, we have the joint distribution of data pair (t, x) as

$$\begin{aligned} & p(t, x \mid \mu, W, \sigma^2) \\ &= p(t \mid x, \mu, W, \sigma^2) \times P(x) \\ &= (2\pi\sigma^2)^{-d/2} \exp\left\{ -\frac{\|t - Wx - \mu\|^2}{2\sigma^2} \right\} \times (2\pi)^{-q/2} \exp\left\{ -\frac{1}{2} x^T x \right\} \end{aligned} \tag{7.56}$$

7.4. PPCA

Suppose we collect a set of observation on $\{t_1, t_2, ..., t_N\}$. To achieve our objective, we maximize the following likelihood function

$$l_C = \prod_{n=1}^{N} \{ p(t_n, x_n \mid W, \mu, \sigma^2) \} = \prod_{n=1}^{N} \{ p(t_{n\_miss}, t_{n\_obse}, x_n \mid W, \mu, \sigma^2) \} \quad (7.57)$$

If t is known, we can optimize this likelihood function directly. If t consisits of two parts: the missing part t_{miss} can be isolated from the known part t_{obse}, we apply the famous EM algorithms.

7.4. PPCA

The parameter μ can be calculated using the average of the observable part t_{obse} and initialize the missing part t_{miss} using μ.

In the E-Step, we estimate the miss part of t_n from our estimation of W and x_n. Suppose $W(k)$ and $E[x_n(k)]$ are the estimation values in the k th iteration, we have

$$t_n(k) = W(k)E[x_n(k)] + \mu \tag{7.58}$$

7.4. PPCA

The expectation of sufficient statistics are

$$E(x_n) = M(k)^{-1} W(k)^T \left(t_n(k) - \mu \right) \tag{7.59}$$

because of the conditional probability distribution function on x_n given t_n:

$$p(x_n \mid t_n, W, \sigma^2, \mu) = N(x_n \mid M^{-1} W^T (t_n - \mu), \sigma^2 M^{-1}) \tag{7.60}$$

where M is defined as $M = W^T W + \sigma^2 I$.

7.4. PPCA

In the M-Step, the expectation of L_C will be maximized with respect to W and σ^2 by the following equations:

$$W(k+1) = S(k)W(k)\left(\sigma(k)^2 I + M(k)^{-1}W(k)^T S(k)W(k)\right)^{-1}$$

$$\sigma(k+1)^2 = \frac{1}{d} tr\left(S(k) - S(k)W(k)M(k)^{-1}W(k+1)^T\right) \qquad (7.61)$$

where

$$S(k) = \frac{1}{N}\sum_{n=1}^{N}\left(t_n(k) - \mu\right)\left(t_n(k) - \mu\right)^T \qquad (7.62)$$

7.4. PPCA

Generalized E-step: For each (possibly incomplete) point t find the unique pair of points x^* and t^* (such that x^* lies in the current principal subspace and t^* lies in the subspace defined by the known information about t) which minimize the norm $\|t^* - Wx^*\|$.

在我们求得最优模型参数的同时，我们也同时补偿了相应的缺失数据。Cheers！☺

PPCA 的最佳隐空间维度可以通过 Bayesian PCA 方法预先测试得到

7.5. References

[1] S. C. Choi, R. Wette, "Maximum Likelihood Estimation of the parameters of the Gamma distribution and their bias," *Technometrics*, 11, no. 4, pp. 683-690, 1969.

[2] E. L. Lehmaan, G. Casella, *Theory of Point Esitmation*, 2nd. ed., Springer, New York, NY, USA, 1998.

[3] G. Casella, R. L. Berger, *Statistical Inference*, 2nd edition, Duxbury and Thomson Learning, 2002.

[4] M. Evans, N. Hastings, B. Peacock, *Statistical Distributions*, 3rd edition, John Wiley & Sons, New York, NY, 2000.

[5] S, Weisberg, *Applied Linear Regression*, 3rd edition, Wiley-Interscience, 2005. http://www.stat.umn.edu/alr/

[6] B. Efron, "Maximum likelihood and decision theory," *Annals of Statistics*, vol. 10, no. 2, pp. 340-356, 1982.

[7] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd edition, Springer, 2009.

[8] H. Hartley, "Maximum likelihood estimation from incomplete data," *Biometrics*, vol. 14, no. 2, pp. 174-194, 1958.

[9] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B, Methodological*, vol. 39, no. 1. pp. 1-38, 1977.

[10] C. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, no. 1, pp. 9-103, March 1983.

[11] J. Ma, L. Xu, M. Jordan, "Asymptotic convergence rate of the EM algorithm for Gaussian mixtures," *Neural Computation*, vol. 12, no. 12, pp. 2881-2907, 2000.

[12] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, 2nd edition, John Wiley & Sons, 2008.

[13] M. R. Gupta, Y. Chen, "Theory and use of the EM algorithm," *Foundations and Trends® in Signal Processing*, vol. 4, no. 3, pp. 223-296, 2010.

[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[15] Jensen, J. L. W. V. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175-193, 1906.

[16] http://en.wikipedia.org/wiki/Expectation-maximization_algorithm

[17] http://en.wikipedia.org/wiki/Gibbs'_inequality

[18] S. Kullback, R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79-86, 1951.

[19] S. Kullback, "Letter to the Editor: The Kullback-Leibler distance," *The American Statistician*, vol. 41, no. 4, pp. 340-341, 1987.

[20] http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

[21] A. Dempster, N. Laird, D. Rubin, "Maximum likelihood from incomplete data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B, Methodological*, vol. 39, no. 1. pp. 1-38, 1977.

[22] S. Roweis, "EM Algorithms for PCA and SPCA," *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*, vol. 10, pp. 626-632, 1998.

[23] M. E. Tipping, C. M. Bishop, "Probabilistic Principal Compnent Analysis," *Journal of the Royal Statistical Society, Series B*, vol. 21, no. 3, pp. 611-622, 1999.