Prediction and Estimation

Anders Hansson and Martin Andersen

Linköping University and Technical Unversity of Denmark

April 18, 2024

Prediction

For a joint pdf $f_{X,Y} : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R}_+$ of two random variables X and Y with marginal pdfs $f_X : \mathbf{R}^m \to \mathbf{R}_+$ and $f_Y : \mathbf{R}^n \to \mathbf{R}_+$ we are given an observation x of X and would like to *predict* a value y for Y.

The conditional pdf for Y given X: $f_{Y|X} : \mathbf{R}^m \times \mathbf{R}^n \to \mathbf{R}_+$:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

is one solution—*Bayesian approach*. We could also maximize w.r.t. y to get a point estimate— *Maximum a Posteriori* (MAP) estimate.

Another possibility is to look for a predictor $g : \mathbf{R}^m \to \mathbf{R}^n$ that solves

$$\text{minimize}_g E\left(Y - g(X)\right)^T \left(Y - g(X)\right)$$

Conditional Expectation

$$\int_{\mathbf{R}^m} f_X(x) \int_{\mathbf{R}^n} f_{Y|X}(y|x) \left(y^T y - 2g^T(x)y + g^T(x)g(x) \right) dydx$$
$$= \int_{\mathbf{R}^m} f_X(x) \Big(E(Y^T Y|X = x) - (E(Y|X = x))^T E(Y|X = x)$$
$$+ (g(x) - E(Y|X = x))^T (g(x) - E(Y|X = x)) \Big) dx$$

Minimum attained for $g(x) = E(Y|X = x)$, i.e. the conditional expectation of Y given $X = x$.

Minimal value of objective function is the conditional covariance of Y.

Normal Distribution Case

Let

$$f_{X,Y}(z) = \frac{1}{\sqrt{(2\pi)^{m+n}\det\Sigma}} e^{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)}$$

where $z = (x, y)$, $\mu = (\mu_x, \mu_y)$ and where

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix}$$

The *Schur complement formula*:

$$\begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix} = \begin{bmatrix} I & 0 \\ \Sigma_{xy}^T\Sigma_x^{-1} & I \end{bmatrix} \begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y - \Sigma_{xy}^T\Sigma_x^{-1}\Sigma_{xy} \end{bmatrix} \begin{bmatrix} I & \Sigma_x^{-1}\Sigma_{xy} \\ 0 & I \end{bmatrix}$$

where

$$\begin{bmatrix} I & \Sigma_x^{-1}\Sigma_{xy} \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & -\Sigma_x^{-1}\Sigma_{xy} \\ 0 & I \end{bmatrix}$$

Factorization of pdf

We have $f_{X,Y}(z) = f_X(x) f_{Y|X}(y|x)$, where

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma_x}} e^{-\frac{1}{2}(x-\mu_x)^T \Sigma_x^{-1}(x-\mu_x)}$$

and where

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_{y|x}}} e^{-\frac{1}{2}(y-\mu_{y|x})^T \Sigma_{y|x}^{-1}(y-\mu_{y|x})}$$

where

$$\mu_{y|x} = \mu_y + \Sigma_{xy}^T \Sigma_x^{-1}(x - \mu_x); \quad \Sigma_{y|x} = \Sigma_y - \Sigma_{xy}^T \Sigma_x^{-1} \Sigma_{xy}$$

Hence $f_{Y|X}(y|x)$ is conditional pdf and the conditional expectation is $\mu_{y|x}$, which is affine in x.

Gaussian Mixture Model

Let

$$f_{X,Y}(z) = \sum_{i=1}^{N} \alpha_j f_{X_i,Y_i}(z)$$

where

$$f_{X_i,Y_i}(z) = \frac{1}{\sqrt{(2\pi)^{m+n} \det \Sigma}} e^{-\frac{1}{2}(z-\mu_i)^T \Sigma_i^{-1}(z-\mu_i)}$$

where $z = (x, y)$, $\mu_i = (\mu_{x,i}, \mu_{y,i})$ and where

$$\Sigma_i = \begin{bmatrix} \Sigma_{x,i} & \Sigma_{xy,i} \\ \Sigma_{xy,i}^T & \Sigma_{y,i} \end{bmatrix}$$

Here $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$.

Marginal and Conditional pdfs

$$f_X(x) = \sum_{i=1}^{N} \alpha_i f_{X_i}(x)$$

where

$$f_{X_i}(x) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} e^{-\frac{1}{2}(x-\mu_{x,i})^T \Sigma_{x,i}^{-1}(x-\mu_{x,i})}$$

and hence

$$f_{Y|X}(y|x) = \frac{\sum_{i=1}^{N} \alpha_j f_{X_i,Y_i}(x,y)}{\sum_{i=1}^{N} \alpha_i f_{X_i}(x)}$$

Optimal Predictor

$$E(Y|X=x) = \frac{\sum_{i=1}^{N} \alpha_i \int_{\mathbf{R}^n} y f_{X_i,Y_i}(x,y)dy}{\sum_{i=1}^{N} \alpha_i f_{X_i}(x)}$$

Since $f_{X_i,Y_i}(x,y) = f_{X_i}(x) f_{Y_i|X_i}(y|x)$ we obtain

$$E(Y|X=x) = \frac{\sum_{i=1}^{N} \alpha_i f_{X_i}(x) \int_{\mathbf{R}^n} y f_{Y_i|X_i}(y|x)dy}{\sum_{i=1}^{N} \alpha_i f_{X_i}(x)} = \frac{\sum_{i=1}^{N} \alpha_i f_{X_i}(x)\mu_i(x)}{\sum_{i=1}^{N} \alpha_i f_{X_i}(x)}$$

where

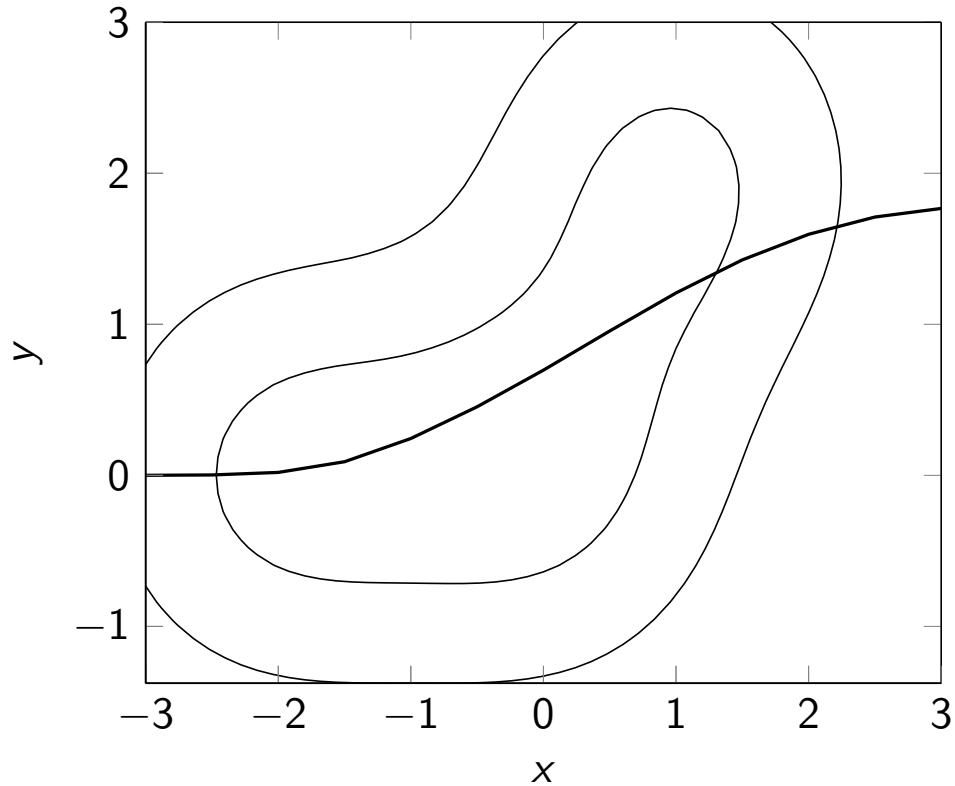
$$\mu_i(x) = \mu_{y,i} + \Sigma_{xy,i}^T \Sigma_{x,i}^{-1}(x - \mu_{x,i})$$

are the linear predictors for Y_i given $X_i = x$.

Example

Two-dimensional Gaussian mixture ($\alpha_i = 1/3$):

$$f_{X_1,Y_1}(x,y) = \mathcal{N}(0, I)$$
$$f_{X_2,Y_2}(x,y) = \mathcal{N}((1,2), I)$$
$$f_{X_3,Y_3}(x,y) = \mathcal{N}((-2,0), I)$$



Affine Predictor

Consider general pdf $f_{X,Y}$ that is not necessarily normal.

Let $m_x = EX$, $m_y = EY$ and $\bar{X} = X - m_x$ and $\bar{Y} = Y - m_y$ and define $J : \mathbf{R}^{m\times n} \times \mathbf{R}^m \to \mathbf{R}_+$ as

$$\begin{aligned}
J(A,b) &= \frac{1}{2} E\left(Y - AX - b\right)^T \left(Y - AX - b\right) \\
&= \frac{1}{2}\,\mathbf{tr}\, E\left(Y - AX - b\right)\left(Y - AX - b\right)^T \\
&= \frac{1}{2}\,\mathbf{tr}\, E\left(\bar{Y} - A\bar{X} + m_y - Am_x - b\right)\left(\bar{Y} - A\bar{X} + m_y - Am_x - b\right)^T \\
&= \frac{1}{2}\,\mathbf{tr}\, E\left(\bar{Y} - A\bar{X}\right)\left(\bar{Y} - A\bar{X}\right)^T \\
&+ \frac{1}{2}\left(m_y - Am_x - b\right)^T \left(m_y - Am_x - b\right)
\end{aligned}$$

where we have used $E\bar{X} = 0$ and $E\bar{Y} = 0$.

Optimal Predictor

With $D_x = E\bar{X}\bar{X}^T$ and $D_{xy} = E\bar{X}\bar{Y}^T$ optimal A and b satisfy

$$\frac{\partial J}{\partial b} = -m_y + Am_x + b = 0$$
$$\frac{\partial J}{\partial A} = -D_{xy}^T + AD_x - m_y m_x^T + bm_x^T + Am_x m_x^T = 0$$

with solution $A = D_{xy}^T D_x^{-1}$ and $b = m_y - D_{xy}^T D_x^{-1} m_x$.

Optimal predictor:

$$Ax + b = m_y + D_{xy}^T D_x^{-1}(x - m_x)$$

which is in agreement with the normal distribution case.

The minimal value of J is $\mathbf{tr}\left(D_y - D_{xy}^T D_x^{-1} D_{xy}\right)$. In general larger that the covariance of Y conditioned on $X = x$.

Affine Predictor for Gaussian Mixture

For general Gaussian mixture:

$$m_x = \sum_{i=1}^{N} \alpha_i \mu_{x,i}; \quad m_y = \sum_{i=1}^{N} \alpha_i \mu_{y,i}$$

Moreover

$$D_x = \sum_{i=1}^{N} \alpha_i \left(\Sigma_{x,i} + \mu_{x,i} \mu_{x,i}^T - m_x m_x^T \right)$$

$$D_{xy} = \sum_{i=1}^{N} \alpha_i (\mu_{x,i} - m_x)(\mu_{y,i} - m_y)^T$$

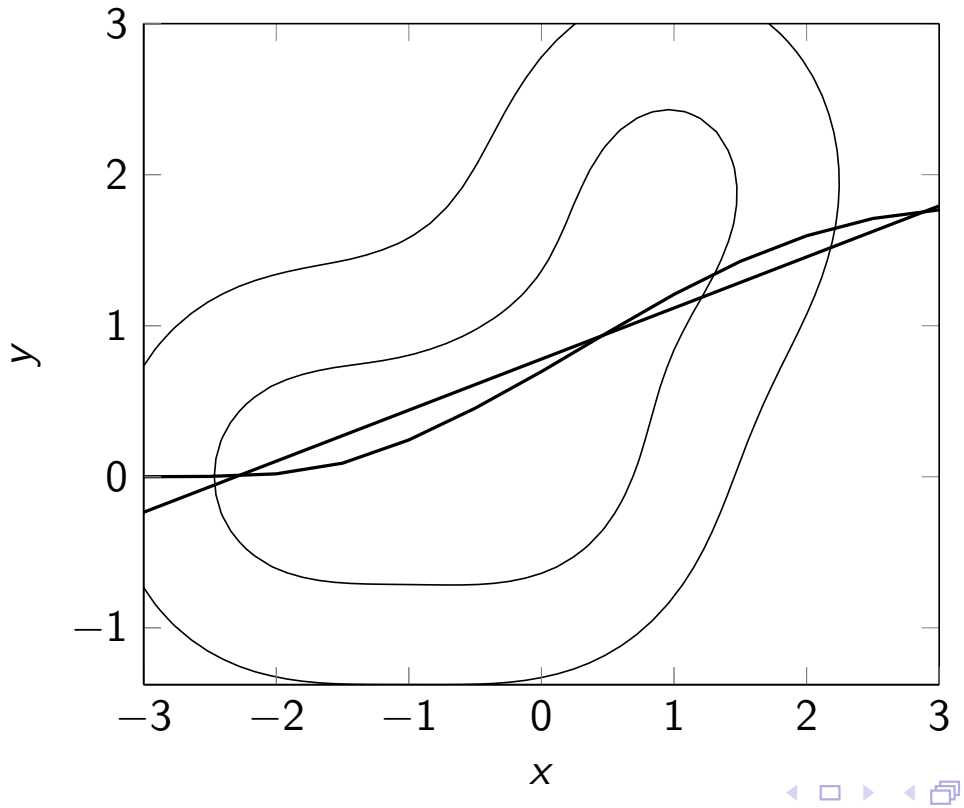
where the latter formula only holds for the case when X_i and Y_i are independent.

Gaussian Mixture Example

For previous example we have $m_x = -1/3$, $m_y = 2/3$, $D_x = 71/27$ and $D_{xy} = 24/27$.

Affine predictor is

$$Ax + b = \frac{2}{3} + \frac{24}{71}\left(x + \frac{1}{3}\right)$$



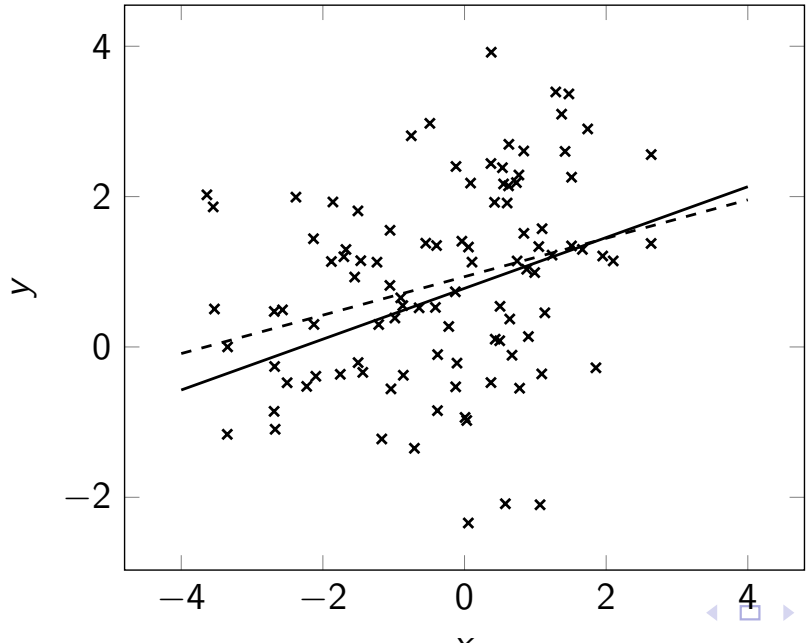
Stochastic Averaging Approximation (SSA)

Replace the expected value with averaging using observations (x_i, y_i) $i \in \mathbf{N}_N$ of the random variables (X, Y), i.e. minimize

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - Ax_i - b)^T(y_i - Ax_i - b)$$

which is a Least Squares (LS) problem.

Solution from affine predictor by replacing true moments with their estimates from the observations (x_i, y_i)



Hidden Markov Model (HMM)

Consider two random processes $X : \Omega \to \mathcal{D}^{\mathbf{Z}_+}$ and $Y : \Omega \to \mathcal{E}^{\mathbf{Z}_+}$ which are correlated with one another.

We assume X is a Markov process and that Y_j given X_j are independent of Y_k given X_k for $j \neq k$.

Smoothing Problem

We want to predict/estimate $\bar{X}_k = (X_0, \ldots, X_k)$ from the observation $\bar{y}_k = (y_0, \ldots, y_k)$ of $\bar{Y}_k = (Y_0, \ldots, Y_k)$ using MAP estimation.

The process X is called the *state* and Y is called the *measurement* or *output*.

Joint pdf and conditional pdf

Start with joint pdf[1] $p_{\bar{X}_k,\bar{Y}_k} : \mathcal{D}^{k+1} \times \mathcal{E}^{k+1} \to \mathbf{R}_+$ for $(\bar{X}_k, \bar{Y}_k)$

Conditional pdf for $\bar{Y}_k$ given $\bar{X}_k$: $p_{\bar{Y}_k|\bar{X}_k} : \mathcal{E}^{k+1} \times \mathcal{D}^{k+1} \to \mathbf{R}_+$ can be expressed

$$p_{\bar{Y}_k|\bar{X}_k}(\bar{y}_k|\bar{x}_k) = \prod_{i=0}^{k} p_{Y_i|X_i}(y_i|x_i)$$

where $p_{Y_i|X_i} : \mathcal{E} \times \mathcal{D} \to \mathbf{R}_+$ are the conditional pdf:s for Y_i given X_i.

Let the marginal probability function or pdf for $\bar{X}_k$ be $p_{\bar{X}_k} : \mathcal{D}^{k+1} \to \mathbf{R}_+$.

Hence

$$p_{\bar{X}_k,\bar{Y}_k}(\bar{x}_k, \bar{y}_k) = p_{\bar{Y}_k|\bar{X}_k}(\bar{y}_k|\bar{x}_k) p_{\bar{X}_k}(\bar{x}_k) = \prod_{i=0}^{k} p_{Y_i|X_i}(y_i|x_i) p_X(\bar{x}_k)$$

[1] We should replace pdf with probability function if the sample space is

From Multiplication Theorem and Markov Property

$$
\begin{aligned}
p_{\bar{X}_k}(\bar{x}_k) &= p_{X_0}(x_0)p_{X_1|X_0}(x_1|x_0)p_{X_2|X_0,X_1}(x_2|x_0,x_1) \\
&\quad \cdots p_{X_k|X_0,\ldots,X_{k-1}}(x_k|x_0,\ldots,x_{k-1}) \\
&= p_{X_0}(x_0)p_{X_1|X_0}(x_1|x_0)p_{X_2|X_1}(x_2|x_1) \\
&\quad \cdots p_{X_k|X_{k-1}}(x_k|x_{k-1})
\end{aligned}
$$

and hence

$$
\begin{aligned}
p_{\bar{X}_k,\bar{Y}_k}(\bar{x}_k,\bar{y}_k) &= p_{Y_k|X_k}(y_k|x_k) \\
&\quad \times p_{X_k|X_{k-1}}(x_k|x_{k-1}) \\
&\quad \times p_{Y_{k-1}|X_{k-1}}(y_{k-1}|x_{k-1}) \\
&\quad \times p_{X_{k-1}|X_{k-2}}(x_{k-1}|x_{k-2}) \\
&\quad \vdots \\
&\quad \times p_{X_1|X_0}(x_1|x_0) \\
&\quad \times p_{Y_0|X_0}(y_0|x_0)p_{X_0}(x_0)
\end{aligned}
$$

Recursive Optimization

$$
\begin{aligned}
\max_{\bar{x}_k} p_{\bar{X}_k,\bar{Y}_k}(\bar{x}_k,\bar{y}_k) = \max_{x_k}\{ & p_{Y_k|X_k}(y_k|x_k) \\
& \times \max_{x_{k-1}}\{p_{X_k|X_{k-1}}(x_k|x_{k-1})p_{Y_{k-1}|X_{k-1}}(y_{k-1}|x_{k-1}) \\
& \times \max_{x_{k-2}}\{p_{X_{k-1}|X_{k-2}}(x_{k-1}|x_{k-2})p_{Y_{k-2}|X_{k-2}}(y_{k-2}|x_{k-2}) \\
& \vdots \\
& \times \max_{x_1}\{p_{X_2|X_1}(x_2|x_1)p_{Y_1|X_1}(y_1|x_1) \\
& \times \max_{x_0}\{p_{X_1|X_0}(x_1|x_0)p_{Y_0|X_0}(y_0|x_0)p_{X_0}(x_0)\}\}\}\}\}
\end{aligned}
$$

Viterbi Algorithm

Let $V_k : \mathcal{D} \to \mathbf{R}_+$ for $0 \leq k \leq N$ be defined via $V_0(x) = p_{Y_0|X_0}(y_0|x)p_{X_0}(x)$ and the recursion

$$V_i(x) = p_{Y_i|X_i}(y_i|x) \max_u p_{X_i|X_{i-1}}(x|u) V_{i-1}(u)$$

for $k = 1, \ldots, k$.

Then

$$\max_{\bar{x}_k} p_{\bar{X}_k, \bar{Y}_k}(\bar{x}_k, \bar{y}_k) = \max_x V_k(x)$$

and the optimal $\bar{x}_k$ is such that x_{i-1} is the maximizing u in iteration i above.

Logarithmic Viterbi Algorithm

Let $J_i : \mathcal{D} \to \mathbf{R}$ for $0 \leq i \leq k$ as $J_i(x) = -\log V_i(x)$.

Then

$$J_i(x) = -\log p_{Y_i|X_i}(y_i|x) + \min_u \{-\log p_{X_i|X_{i-1}}(x|u) + J_{i-1}(u)\}$$

with initial value $J_0(x) = -\log p_{Y_0|X_0}(y_0|x) - \log p_{X_0}(x)$.

Often better numerical properties.

Dynamical State Equation

A typical example of an HMM for $\mathcal{D} = \mathbf{R}^n$ and $\mathcal{E} = \mathbf{R}^p$:

$$\begin{aligned} X_{k+1} &= F_k(X_k, V_k) \\ Y_k &= G_k(X_k, E_k) \end{aligned} \tag{1}$$

where $F_k : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}^n$, $G_k : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}^p$, and where E_k are i.i.d. p-dimensional random vectors and V_k are i.i.d. n-dimensional random vectors, and where $X_0 \in \mathbf{R}^n$ is a random vector with known distributions. Moreover, X_0 is independent of E_k and V_k for all $k \geq 0$. and E_k and V_k are independent for all $k \geq 0$.

Linear Dynamics

Consider $F(x, v) = Ax + v$ and $G(x, e) = Cx + e$, where $A \in \mathbf{R}^{n \times n}$, and $C \in \mathbf{R}^{p \times n}$.

Assume X_0, V_k and E_k all have Gaussian distributions with expectations $\bar{x}$, 0 and 0, respectively, and covariances R_0, R_1 and R_2, respectively.

Then

$$
\begin{aligned}
p_{X_0}(x_0) &= \mathcal{N}(x_0, \bar{x}_0, R_0) \\
p_{X_k|X_{k-1}}(x_k|x_{k-1}) &= \mathcal{N}(x_k, Ax_{k-1}, R_1) \\
p_{Y_k|X_k}(y_k|x_k) &= \mathcal{N}(y_k, Cx_k, R_2)
\end{aligned}
$$

Logarithmic Viterbi Recursion

$$J_0(x) = \frac{1}{2}(x - \bar{x}_0)^T R_0^{-1}(x - \bar{x}_0) + \frac{1}{2}(y_0 - Cx)^T R_2^{-1}(y_0 - Cx) \tag{2}$$

$$J_i(x) = \frac{1}{2}(y_k - Cx)^T R_2^{-1}(y_k - Cx) \tag{3}$$

$$+ \min_u \left\{ \frac{1}{2}(x - Au)^T R_1^{-1}(x - Au) + J_{i-1}(u) \right\} \tag{4}$$

for $i = 1, \ldots, k$ modulo constant terms.

Ansatz

$$J_i(x) = \frac{1}{2}x^T P_i x + q_i^T x + r_i$$

for some $P_i \in \mathbf{S}_+^n$, $q_i \in \mathbf{R}^n$ and $r_i \in \mathbf{R}$.

For $i = 0$ this holds with

$$P_0 = R_0^{-1} + C^T R_2^{-1} C; \quad q_0 = R_0^{-1}\bar{x}_0 + C^T R_2^{-1} y_0$$

The right hand side of (4) minimize when gradient is zero:

$$-A^T R_1^{-1}(x - Au) + P_{i-1}u + q_{i-1} = 0$$

$$\Longleftrightarrow$$

$$u = G_{i-1}^{-1}\left(A^T R_1^{-1} x - q_{i-1}\right)$$

where $G_i = A^T R_1^{-1} A + P_i$.

Sufficent Condition for Recursion to Hold

$$P_i = C^T R_2^{-1} C + Y_i^f$$
$$q_i = -C^T R_2^{-1} y_i - Y_k^f A P_{i-1}^{-1} q_{i-1}$$

where

$$Y_i^f = \left(R_1 + A P_{i-1}^{-1} A^T \right)^{-1}$$

Optimal Estimates

$$x_k = -P_k^{-1} q_k = P_k^{-1} \left(Y_k^f A P_{k-1}^{-1} q_{k-1} + C^T R_2^{-1} y_k \right)$$

We can obtain the solution for the problem ending at $k+1$ from the solution for the problem ending at k with just one more step in the recursion.

We may also use $u = x_{i-1}$ to obtain the estimates of x_i for $0 \leq i \leq k-1$, which are the so-called smoothed estimates:

$$x_{i-1} = G_{i-1}^{-1} \left(A^T R_1^{-1} x_i - q_{i-1} \right)$$

From matrix inversion lemma as

$$G_{i-1}^{-1} = P_{i-1}^{-1} - P_{i-1}^{-1} A^T \left(R_1 + A P_{i-1}^{-1} A^T \right)^{-1} P_{i-1}^{-1}$$

The smoothed estimates will be different for different values of k, i.e. we cannot find the smoothed solution for the problem ending at $k+1$ from the solution of the problem ending at k without re-running the above backward recursion.

Graphical Models

Consider a graph $G = (V, E)$, where $V = \mathbf{N}_n$ is the set of vertices and where $E \subset V \times V$ is a set of undirected edges connecting vertices in V.

We let the elements of V index components of an n-dimensional random variable.

We define similarly as before a distribution function by maximizing entropy under moment constraints. However, we only specify second order moments for combinations of components belonging to E.

Graphical Ising Distribution

When we do this for an n-dimensional Ising distribution we obtain

$$p(x) = \exp\left(\sum_{k\in\mathbf{N}_n} \lambda_k x_k + \sum_{(i,j)\in E} \Lambda_{i,j} x_i x_j - A\right)$$

where

$$A(\lambda,\Lambda) = \ln \sum_{x\in\mathcal{D}_x} \exp\left(\sum_{k\in\mathbf{N}_n} \lambda_k x_k + \sum_{(i,j)\in E} \Lambda_{i,j} x_i x_j\right)$$

We notice that we just do not specify all the entries of neither Λ nor M.

Matching conditions

If we let $\Lambda_{i,j} = 0$ for $(i,j) \notin E$, then we may write the distribution also in this case using $\mathbf{tr}\,\Lambda x x^T$. The Lagrange dual function will therefore look the same and so will the optimality conditions, except for the fact that some of them are void, i.e. we only have

$$\frac{\partial h}{\partial \lambda} = \frac{\partial A}{\partial \lambda} - m = 0$$
$$\frac{\partial h}{\partial \Lambda_{i,j}} = \frac{\partial A}{\partial \Lambda_{i,j}} - M_{i,j} = 0, \quad (i,j) \in E$$

Also for this case it is in general difficult to solve the optimality conditions.

Graphical Normal Distribution

Similarly, when we derive the normal distribution the only difference is that Λ is such that $\Lambda_{i,j} = 0$ for $(i,j) \notin E$. We do not need to specify $M_{i,j}$ for $(i,j) \notin E$.

The Lagrange dual function will look the same. We first minimize it with respect to λ, which results in $\lambda = -\Lambda m$. We then back substitute in order to obtain the function $f : \mathbf{S}^n \to \mathbf{R}$ given by

$$f(\Lambda) = -\frac{1}{2} \ln \frac{\det \Lambda}{(2\pi)^n} + \frac{1}{2} \mathbf{tr}\, \Lambda\Sigma$$

where $\Sigma = M - mm^T$.

We need to minimize f subject to the constraint that $\Lambda_{i,j} = 0$ for $(i,j) \notin E$ in order to express Λ in terms of the moments.

This is a convex optimization problem with a linear constraint on Λ.

Further Analysis

Lagrangian $\mathcal{L} : \mathbf{S}^n \times \mathbf{S}^n$ defined by

$$\mathcal{L}(\Lambda, \Gamma) = -\frac{1}{2} \ln \frac{\det \Lambda}{(2\pi)^n} + \frac{1}{2} \mathbf{tr}\, \Lambda\Sigma + \frac{1}{2} \mathbf{tr}\, \Gamma\Lambda$$

where $\Gamma_{i,j} = 0$ for $(i,j) \in E$.

Optimality conditions

$$\frac{\partial \mathcal{L}}{\partial \Lambda} = -\frac{1}{2}\Lambda^{-1} + \frac{1}{2}\Sigma + \frac{1}{2}\Gamma = 0$$

together with $\Lambda_{i,j} = 0$ for $(i,j) \notin E$ and $\Gamma_{i,j} = 0$ for $(i,j) \in E$.

Positive Definite Completion

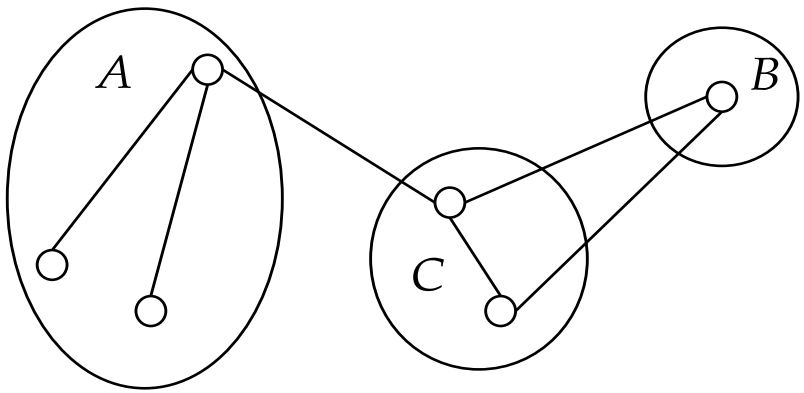
We realize that we have full freedom in selecting the entries of $\Sigma + \Gamma$ for indexes $(i, j) \notin E$ by choosing Γ. Hence it does not matter that we did not specify these entries for M. We may take them equal to any value. Because of this we let $\Sigma_{i,j} = 0$ for $(i, j) \notin E$.

However, $\Sigma + \Gamma$ will be the covariance matrix of the normal distribution, and hence has to be positive definite.

In summary we want to find Γ to complete the covariance matrix in such a way that it is positive definite and such that its inverse has zeros for entries not in the set E.

This is called the *postive definite completion* of Σ.

Conditional Independence



Consider

$$p(x) = \sqrt{\frac{\det \Lambda}{\pi^n}} e^{-(x-m)^T \Lambda (x-m)}$$

Assume $V = A \cup B \cup C$, and order such that $x = (x_A, x_B, x_C)$.

Implies Zero Structure

$$\Lambda = \begin{bmatrix} * & 0 & * \\ 0 & * & * \\ * & * & * \end{bmatrix}$$

and

$$p(x) = p_{A,B|C}(x)p_C(x_C) = p_{A|C}(x_A, x_C)p_{B|C}(x_B, x_C)p_C(x_C)$$

Markov Random Field

In general p can be factorized as

$$p(x) = \prod_{C \in \mathcal{C}} f_C(x_C)$$

for some functions $f_C : \mathbf{R}^{|C|} \to \mathbf{R}_+$, where $\mathcal{C}$ is the set of all *cliques* of G, i.e. the set of all complete subgraphs of G.

The above Markov property holds for general graphical models defined on undirected graphs, and specifically also for the Ising model.

Markov processes a special case.

Maximum Likelihood Estimation

For a probability function $p \in \mathcal{D}_n$ that depends on a parameter $\lambda \in \mathbf{R}$ we may define the likelihood function $\ell : \mathbf{R}^N \times \mathbf{R} \to [0, 1]$ based on N samples f_{k_i}, $i \in \mathbf{N}_N$ of the random variable $X : \mathbf{N}_n \to \{f_1, \ldots, f_n\} \subset \mathbf{R}^n$ as

$$\ell(f_{k_1}, \ldots, f_{k_N}; \lambda) = \prod_{i=1}^{N} p_{k_i}(\lambda)$$

Then the estimate of λ is obtained by maximizing ℓ.

For Categorial distribution:

$$\ln \ell(f_{k_1}, \ldots, f_{k_N}; \lambda) = -\sum_{i=1}^{N} \lambda f_{k_i} - N \ln \Phi(\lambda)$$

With $b = \frac{1}{N} \sum_{i=1}^{n} f_{k_i}$ minimizing "Lagrange dual function" $h(\lambda)$ is equivalent to maximizing $\ln \ell(f_{k_1}, \ldots, f_{k_N}; \lambda)$.

Ising Distribution

For Ising distribution $\ell : \{0,1\}^{mN} \times \mathbf{R}^m \times \mathbf{S}^m \to [0,1]$ based on N samples $x_i \in \{0,1\}^m$, $i \in \mathbf{N}_N$:

$$\ell(x_1, \ldots, x_N; \lambda, \Lambda) = \prod_{i=1}^{N} p(x_i)$$

we have

$$\ln \ell(x_1, \ldots, x_N; \lambda, \Lambda) = \lambda^T \sum_{i=1}^{N} x_i + \mathbf{tr}\, \Lambda \sum_{i=1}^{N} x_i x_i^T - NA(\lambda, \Lambda)$$

With $m = \frac{1}{N} \sum_{i=1}^{N} x_i$ and $M = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$ minimizing "Lagrange Dual function" $h(\lambda, \Lambda)$ is equivalent to maximizing the likelihood function.

Also true for graphical Ising model.

Normal Distribution

For the normal distribution with $\theta = (\lambda, \Lambda) \in \mathbf{R}^n \times \mathbf{S}^n$:

$$\ln \ell(x_1, \ldots, x_N; \theta) = -\frac{1}{2} \mathbf{tr}\, \Lambda \left(\sum_{i=1}^{N} x_i x_i^T \right) - \lambda^T \sum_{i=1}^{N} x_i - \frac{N}{2} \lambda^T \Lambda^{-1} \lambda + \frac{N}{2} \ln \frac{\det \Lambda}{(2\pi)^n}$$

With

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i; \qquad M = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$$

minimizing "Lagrange dual function" $h(\Lambda)$ is equivalent to maximizing the likelihood function.

The solution has already been computed and with $\Sigma = M - mm^T$ we have $\Lambda = \Sigma^{-1}$ and $\lambda = -\Sigma^{-1} m$.

Similar results for graphical normal distribution

Generalizations

With prior information like

$$B_l \preceq \Lambda \preceq B_u$$

with $B_l, B_u \in \mathbf{S}^n_{++}$ we have a convex constraint which can be incorporated when we minimize $h(\lambda, \Lambda)$.

Also an upper bound $\kappa_{\max}$ on the condition number of Λ can be incorporated by noting that it is equivalent to the existence of $u > 0$ such that $uI \preceq \Lambda \preceq \kappa_{\max} uI$.

Prior information can be included:

$$m = \alpha m_0 + (1 - \alpha)\frac{1}{N}\sum_{i=1}^{N} x_i$$

$$M = \beta M_0 + (1 - \beta)\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$$

with $\alpha, \beta \in [0, 1]$.