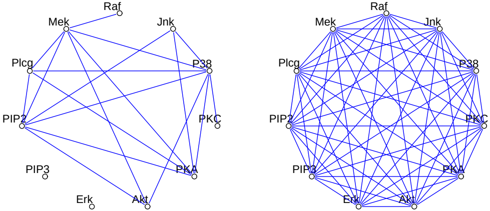
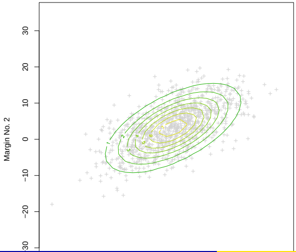
Undirected (Pairwise, Continuous) Graphical Models

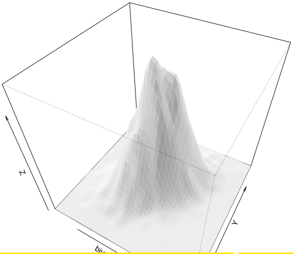
- No specific goal variable
- Any variable may be taken as the goal
- the **generative model** represents the full probability distribution $P(X)$.
- Missing edges represent conditional independence of the variables.



Gaussian Graphical Models

- **Multivariate Gaussian Distribution** on variables $X = (X_1, \ldots, X_p)$
- $\phi(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)}$
  - $|.|$ is the determinant. we denote p the number of components in $\mathbf{x}$. Then $|2\pi\Sigma| = (2\pi)^p|\Sigma|$.
- If Σ is not invertible it has dependent columns. It means that the variables $\mathbf{x}_j$ are lineary dependent.
  - If the **rank** of Σ is ℓ then there exists a matrix A and a vector ν so:
  - $x = Az + \nu$ for new coordinates z with ℓ dimensions
  - We just consider the new coordinates and assume Σ has a full rank.
- Applications: genomics and proteomics, Cyclomeric dataset in ESLII.





Data: carcass

Data: carcass #Source: Soren Hojsgaard, David Edwards, Steffen Lauritzen: *Graphical Models with R*, Springer.

	mean.
Fat11	16.00
Meat11	52.00
Fat12	14.00
Meat12	52.00
Fat13	13.00
Meat13	56.00
LeanMeat	59.00

Σ	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	LeanMeat
Fat11	11.34	0.74	8.42	2.06	7.66	-0.76	-9.08
Meat11	0.74	32.97	0.67	35.94	2.01	31.97	5.33
Fat12	8.42	0.67	8.91	0.31	6.84	-0.60	-7.95
Meat12	2.06	35.94	0.31	51.79	2.18	41.47	6.03
Fat13	7.66	2.01	6.84	2.18	7.62	0.38	-6.93
Meat13	-0.76	31.97	-0.60	41.47	0.38	41.44	7.23
LeanMeat	-9.08	5.33	-7.95	6.03	-6.93	7.23	12.90

Concentration matrix

- **Concentration (Precision, koncentrační) matrix**

$$K = \Sigma^{-1}$$

Lemma

For $u \neq v$, $k_{uv} = 0$ if and only if y_u and y_v are conditionally independent given all other variables.

k*100	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	LeanMeat
Fat11	44	3	-20	-7	-16	4	10
Meat11	3	16	-3	-6	-6	-6	-3
Fat12	-20	-3	54	6	-21	-5	9
Meat12	-7	-6	6	14	-1	-9	-0
Fat13	-16	-6	-21	-1	56	3	7
Meat13	4	-6	-5	-9	3	16	-1
LeanMeat	10	-3	9	-0	7	-1	26

Partial correlation matrix

Definition (Partial correlation matrix)

Partial correlation matrix is defined from K by

$$\rho_{uv|V\setminus\{uv\}} = \frac{-k_{uv}}{\sqrt{k_{uu}k_{vv}}}.$$

Lemma

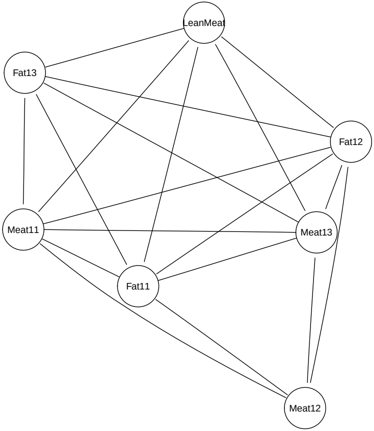
In contrast to concentrations, the partial correlations are invariant under a change of scale and origin in the sense that if $X_j^* = a_j X_j + b_j$, $j = 1, \dots, p$ then $a_v a_u k_{uv}^* = k_{uv}$ and $\rho_{uv|V\setminus\{uv\}}^* = \rho_{uv|V\setminus\{uv\}}$.

$\rho * 100$	Fat11	Meat11	Fat12	Meat12	Fat13	Meat13	LeanMeat
Fat11	-	-11	41	30	32	-16	-29
Meat11	-11	-	9	41	19	35	16
Fat12	41	9	-	-24	38	18	-24
Meat12	30	41	-24	-	2	61	2
Fat13	32	19	38	2	-	-9	-18
Meat13	-16	35	18	61	-9	-	7
LeanMeat	-29	16	-24	2	-18	7	-

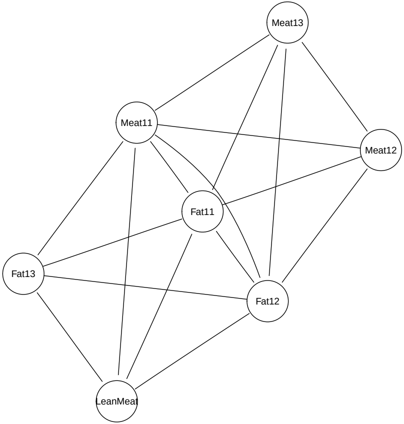
Models

The simplest model just removes edges with small $|\rho_{uv|V\setminus\{uv\}}|$. More advanced AIC, BIC criteria will be introduced later.

AIC



BIC



Undirected Graphical Models and Their Properties

Definition (Undirected Graphical Model, Markov Graph)

An **Undirected Graphical Model** (Markov graph, Markov network) is a graph $\mathcal{G} = (V, E)$, where nodes V represent random variables and the absence of an edge (A, B) denoted $A \perp\!\!\!\perp B$ implies that the corresponding random variables are conditionally independent given the rest.

$$A \perp\!\!\!\perp B \Longrightarrow A \perp B | V \setminus \{A, B\}. \tag{4}$$

(4) is known as the **pairwise Markov independencies** of $\mathcal{G}$.

Definition (Separators)

- If A, B and C are subgraphs, then C **is said to separate** A **and** B if every path between A and B intersects a node in C.
- C is called a **separator**.

- Separators break the graph into conditionally independent pieces.

Markov Properties

Definition (Global Markov Property)

A Markov graph $\mathcal{G}$ fulfills a **Global Markov Property** (5) iff for any subgraphs A, B and C holds:

- if C separates A and B then $A \perp B | C$, that is

$$A \perp\!\!\!\perp B | C \Longrightarrow A \perp B | C. \tag{5}$$

Theorem

The pairwise and global Markov properties of a graph are equivalent for graphs with positive distributions.

- Gaussian distribution is always positive.
- We may infer global independence relations from simple pairwise properties.
- The global Markov property allows us to decompose graphs into smaller more manageable pieces.

Markov Random Fields (Markovská náhodná pole)

- A probability density function f over a Markov graph $\mathcal{G}$ with the set of maximal cliques $\{C_1, \ldots, C_k\}$ can be represented as

$$f(x) = \prod_{i=1,\ldots,k} \psi_i(x_{C_i}) = \psi_1(x_{C_1}) \cdot \ldots \cdot \psi_k(x_{C_k}) \tag{6}$$

- where ψ_i are positive functions called **clique potentials**.
- they capture the dependence in X_{C_i} by scoring certain instances x_{C_i} higher than others.
- with the **normalizing constant** (partition function) Z

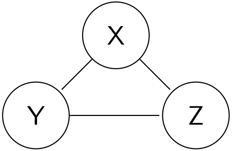
$$Z = \int_X exp\left(\sum_{i=1,\ldots,k} \log g_i(y_{C_i}) \right).$$

- For Markov networks with positive distributions the probability density function (6) implies a graph with independence properties defined by the cliques in the product.

Pairwise Markov Graphs

- A graphical model does not always uniquely specify the higher–order dependence structure of ta joint probability distribution.

$$\begin{aligned} f^{(2)}(x,y,z) &= \frac{1}{Z}\psi(x,y)\psi(x,z)\psi(y,z) \\ f^{(3)}(x,y,z) &= \frac{1}{Z}\psi(x,y,z) \end{aligned}$$



- For Gaussian distribution, parwise interactions fully specify the model.
- We focus on **pairwise Markov Graphs**
  - where at most second order interactions are represented (like $f^{(2)}$).

Undirected Gaussian graphical model

Definition (Undirected Gaussian graphical model)

An **undirected Gaussian graphical model** is represented by an undirected graph $\mathcal{G} = (X, E)$, $X = \{X_1, \dots, X_p\}$ represent the set of variables and E is a set of undirected edges.
When a random vector $\mathbf{x}$ follows a Gaussian distribution $N_p(\mu, \Sigma)$, the graph G represents the model where $K = \Sigma^{-1}$ is a positive definite matrix with $k_{u,v} = 0$ whenever there is no edge between vertices u, v in G.
This graph is called the **dependence graph** of the model.

Lemma

For any non adjacent vertices $u, v \in \mathcal{G}$ it holds: $u \perp\!\!\!\perp v | \mathbf{X} \setminus \{u, v\}$.

Definition (Generating class)

Let $\mathcal{C} = \{C_1, \dots, C_k\}$ be the set of cliques of the dependence graph $\mathcal{G}$. A set of functions $g_1(), g_2(), \dots, g_k()$ defined on $g_i(\mathbf{x}_{C_i})$ is called a **generating class** for the distribution

$$f(\mathbf{x}) = \prod_{i=1,\dots,k} g_i(\mathbf{x}_{C_i}).$$

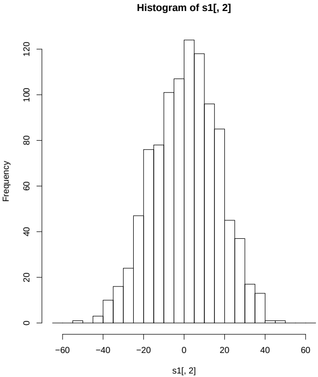
Marginalization

- We have $\frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)}$
- We want the distribution over variables $\{x_3, x_5, x_7\} \subset \{x_1, \ldots, x_p\}$

Marginal of a Gaussian Distribution

The marginal of a Gaussian distribution is calculated by removing appropriate dimensions from the mean and covariance matrix.

- $\mu_{3,5,7} = (\mu_3, \mu_5, \mu_7)$ and $\Sigma_{3,5,7} = \begin{bmatrix} \Sigma_{33} & \Sigma_{35} & \Sigma_{37} \\ \Sigma_{53} & \Sigma_{55} & \Sigma_{57} \\ \Sigma_{73} & \Sigma_{75} & \Sigma_{77} \end{bmatrix}$
- $\phi_{x_3,x_5,x_7} = \frac{1}{\sqrt{|2\pi\Sigma_{3,5,7}|}} e^{-\frac{1}{2}(x_{3,5,7}-\mu_{3,5,7})\Sigma_{3,5,7}^{-1}(x_{3,5,7}-\mu_{3,5,7})}$



Conditioning

- We ame for $\phi(A|B)$ where
  - $A \subset \{x_1, \dots, x_p\}$ having q elements,
  - the rest $B = \{x_1, \dots, x_p\} \setminus A$ has $(p-q)$ elements.
- We rearrange the rows and columns to have A together. Then we get $x = \begin{bmatrix} x_A \\ x_B \end{bmatrix}$ (one column), $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$ (one column), $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$ with dimensions $\begin{bmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{bmatrix}$.

Conditional Gaussian

The parameters of the conditional Gaussian distribution $\phi(A|B=b) = N(\mu_{A|B=b}, \Sigma_{A|B=b})$ are:

$$\begin{aligned} \mu_{A|B=b} &= \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b - \mu_B) \\ \Sigma_{A|B=b} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}. \end{aligned}$$

Covariance matrix differs but does not depend on the observation b. It depends on the fact B was observed.

Conditional Gaussian Example

- $\mu^T = (1, 2, 3, 4)$
- $\Sigma = \begin{bmatrix} 10 & 1 & 5 & 4 \\ 1 & 10 & 2 & 6 \\ 5 & 2 & 10 & 3 \\ 4 & 6 & 3 & 10 \end{bmatrix}$
- We observed (X_3, X_4) to be $(2.8, 4.1)$
- We ask for $\phi(A|B) = \phi(\{X_1, X_2\}|\{X_3, X_4\})$
- $\Sigma_{AB} = \begin{bmatrix} 5 & 4 \\ 2 & 6 \end{bmatrix}$
- $\Sigma_{BB} = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$
- $\Sigma_{BB}^{-1} \doteq \begin{bmatrix} 0.11 & -0.033 \\ -0.033 & 0.11 \end{bmatrix}$
- $\Sigma_{AB}\Sigma_{BB}^{-1} \doteq \begin{bmatrix} 0.484 & 0.055 \\ 0.242 & 0.528 \end{bmatrix}$
- $\mu_{A|B=b} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b - \mu_B)$
- $\mu_{A|B} \doteq \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.484 & 0.055 \\ 0.242 & 0.528 \end{bmatrix} \begin{bmatrix} (2.8 - 3) \\ (4.1 - 4) \end{bmatrix}$
- $\mu_{A|B} \doteq \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} -0.0913 \\ 0.0044 \end{bmatrix} = \begin{bmatrix} 0.9087 \\ 2.0044 \end{bmatrix}$
- $\Sigma_{A|B=b} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$
- $\Sigma_{A|B=b} \doteq \begin{bmatrix} 10 & 1 \\ 1 & 10 \end{bmatrix} + \begin{bmatrix} 2.530 & 2.266 \\ 2.266 & 4.136 \end{bmatrix}$
- $\Sigma_{A|B=b} \doteq \begin{bmatrix} 12.530 & 3.266 \\ 3.266 & 14.136 \end{bmatrix}$

Partition Matrix Inverse Properties

$$\begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix} \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} = \begin{pmatrix} I_{AA} & \mathbf{0} \\ \mathbf{0} & I_{BB} \end{pmatrix}$$

- Top right:

$$\begin{aligned} K_{AA}\Sigma_{AB} &+ K_{AB}\Sigma_{BB} = \mathbf{0} \\ -K_{AA}\Sigma_{AB}\Sigma_{BB}^{-1} &= K_{AB} \qquad (7) \\ \Sigma_{AB}\Sigma_{BB}^{-1} &= -K_{AA}^{-1}K_{AB}. \qquad (8) \end{aligned}$$

- and top left, substitute (7):

$$\begin{aligned} K_{AA}\Sigma_{AA} &+ K_{AB}\Sigma_{BA} = I_{AA} \\ K_{AA}\Sigma_{AA} &+ (-K_{AA}\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}) = I_{AA} \\ K_{AA}^{-1} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \qquad (9) \end{aligned}$$

Regression Coefficients

$$\begin{aligned}\mu_{A|B=b} &= \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b-\mu_B)\\ \Sigma_{A|B=b} &= \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\end{aligned}$$

- Consider x_1 to be a linear function of others with the noise $\epsilon_1 \sim N(0, \sigma_1^2)$:

$$x_{1|2\dots p} = a_1 + \beta_{12}x_2 + \beta_{13}x_3 + \dots + \beta_{1p}x_p + \epsilon_1$$

- Set A the first dimension, B the remaining $(p-1) \times (p-1)$ matrix:

$$x_{1|B=(x_2,\dots,x_p)^T} = \mu_{A|B} + \Sigma_{AB}\Sigma_{BB}^{-1}(\begin{bmatrix} x_2 \\ \dots \\ x_p \end{bmatrix} - \mu_B) + \epsilon$$

- Recall (8): $\Sigma_{AB}\Sigma_{BB}^{-1} = -K_{AA}^{-1}K_{AB}$
- then $\sigma_1^2 = \frac{1}{k_{11}}$ with coefficients β

$$(\beta_{12}, \dots, \beta_{1d}) = -\frac{(k_{12}, \dots, k_{1p})}{k_{11}}.$$

Parameter Learning (may be omitted)

- Let us have the data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ over variables $\mathbf{x} \sim N_p(\mu, \Sigma)$
- $S = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{x})^T$ is the empirical covariance matrix.
- Our model is represented by the concentration matrix $\Theta = \Sigma^{-1}$ and mean μ.
- Log-likelihood of the data is

$$logL(\Theta, \mu) = \frac{N}{2} logdet(\Theta) - \frac{N}{2} tr(\Theta S) - \frac{N}{2} (\bar{x} - \mu)^T \Theta (\bar{x} - \mu).$$

- for a fixed Θ is the maximum for μ: $\mu = \bar{x}$ and the last term is 0. We get
- $logL(\Theta, \mu) \propto logdet(\Theta) - tr(\Theta S)$
- where $tr(\Theta S) = \sum_u \sum_v s_{uv} k_{uv}$, therefore only s_{uv} corresponding to non-zero k_{uv} are considered by the sum.
- We replace the equality conditions by Lagrange multiplyers: $\ell_C(\Theta) = logdet(\Theta) - tr(\Theta S) - \sum_{(j,k) \notin E} \gamma_{jk} \theta_{jk}$
- We maximize. The derivative Θ should be zero (Γ is a matrix with non-zero for missing edges):

$$\Theta^{-1} - S - \Gamma = 0$$

Towards the Algorithm (may be omitted)

- We iterate one row/column after another.
- We start with the sample covariance matrix

$$W_0 \leftarrow S$$

- We derive the formula for the last row/column: the derivative

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} - \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix} - \begin{pmatrix} \Gamma_{11} & \gamma_{12} \\ \gamma_{12}^T & \gamma_{22} \end{pmatrix} = 0$$

- The upper right block can be written as $w_{12} - s_{12} - \gamma_{12} = 0$.
- W is inverse of Θ

$$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0^T & 1 \end{pmatrix}$$

- therefore the last column without last row is:

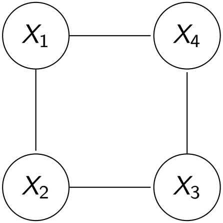
$$w_{12} = -W_{11}\theta_{12}/\theta_{22} = W_{11}\beta$$

- Substitute into the derivative $W_{11}\beta - s_{12} - \gamma_{12} = 0$
- we solve for the rows with zero γ: $\hat{\beta}^* = (W_{11}^*)^{-1} s_{12}^*$.
- The diagonal θ_{22} is (1 bottom right): $\frac{1}{\theta_{22}} = w_{22} - w_{12}^T\beta$.

Estimation of an Undirected Graphical Model Parameters

```
 1: procedure GRAPHICAL REGRESSION:( $S$ sample covariance )
 2:     $W \leftarrow S$ initialize
 3:     repeat
 4:         for $j = 1, 2, \ldots, p$ do
 5:             Partition $W$; $j$th row and column, $W_{11}$ the rest
 6:             solve $W_{11}^{*}\beta^{*} - s_{12}^{*} = 0$ for reduced system
 7:             $\hat{\beta} \leftarrow \hat{\beta}^{*}$ by padding with zeros
 8:             update $w_{12} \leftarrow W_{11}\hat{\beta}$
 9:         end for
10:     until convergence
11:     for $j = 1, 2, \ldots, p$ do
12:         lines 5:-8: above and
13:         solve $\hat{\theta}_{22} \leftarrow \frac{1}{w_{22} - w_{12}^{T}\hat{\beta}}$
14:         solve $\hat{\theta}_{12} \leftarrow -\hat{\beta} \cdot \hat{\theta}_{22}$
15:     end for
16: end procedure
```

Example



$$W_0 = S = \begin{bmatrix} 10.00 & 1.00 & 5.00 & 4.00 \\ 1.00 & 10.00 & 2.00 & 6.00 \\ 5.00 & 2.00 & 10.00 & 3.00 \\ 4.00 & 6.00 & 3.00 & 10.00 \end{bmatrix}$$

$$\begin{aligned} W_{11} &= \begin{bmatrix} 10.00 & 2.00 & 6.00 \\ 2.00 & 10.00 & 3.00 \\ 6.00 & 3.00 & 10.00 \end{bmatrix} \\ W_{11}^{*} &= \begin{bmatrix} 10.00 & 6.00 \\ 6.00 & 10.00 \end{bmatrix} \\ W_{11}^{*,-1} &= \begin{bmatrix} 0.156 & -0.094 \\ -0.094 & 0.156 \end{bmatrix} \\ \beta^{*} &= [-0.22, 0.53]^T \\ \beta &= [-0.22, 0, 0.53]^T \\ w_{12} &\leftarrow [1.00, 1.16, 4.00]^T \end{aligned}$$

$$\begin{aligned} W_{22} &= \begin{bmatrix} 10.00 & 1.16 & 4.00 \\ 1.16 & 10.00 & 3.00 \\ 4.00 & 3.00 & 10.00 \end{bmatrix} \\ W_{22}^{*} &= \begin{bmatrix} 10.00 & 1.16 \\ 1.16 & 10.00 \end{bmatrix} \\ W_{22}^{*,-1} &= \begin{bmatrix} 0.101 & -0.012 \\ -0.012 & 0.101 \end{bmatrix} \\ \beta 2^{*} &= [0.08, 0.19]^T \\ \beta 2 &= [0.08, 0.19, 0]^T \\ w_{2r} &\leftarrow [1.00, 2, 0.88]^T \end{aligned}$$

Structure Learning

- We add a lasso penalty $||\Theta||_1$ which denotes the L_1 norm
  - the sum of the absolute values of the elements of Θ and we ignore the diagonal.
  - The negative penalized log-likelihood is a convex function of Θ.
- we maximize penalized log-likelihood

$$logdet(\Theta) - tr(\Theta S) - \lambda ||\Theta||_1 \tag{10}$$

- the gradient equation is now

$$\Theta^{-1} - S - \lambda Sign(\Theta) = 0 \tag{11}$$

  - sub-gradient notation
  - $Sign(\theta_{jk}) = sign(\theta_{jk})$ for $\theta_{jk} \neq 0$
  - $Sign(\theta_{jk}) \in [-1, 1]$ for $\theta_{jk} = 0$
- the update for the first row and column will be

$$W_{11}\beta - s_{12} + \lambda Sign(\beta) = 0 \tag{12}$$

  - since β and θ_{12} have opposite signs.

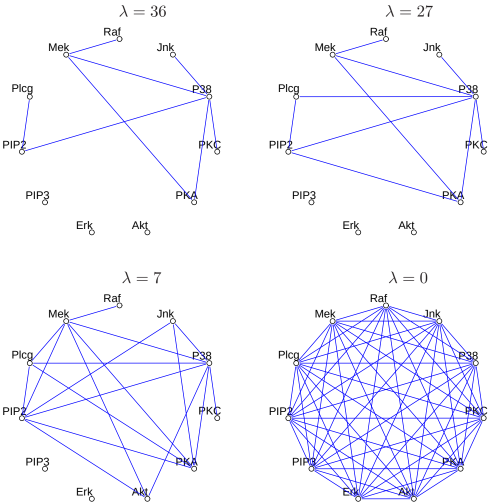
Graphical Lasso

```
procedure GRAPHICAL LASSO:( $S$ sample covariance,$\lambda$ penalty )
    $W \leftarrow S + \lambda I$ initialize
    repeat
        for  $j = 1, 2, \ldots, p$ do
            Partition $W$; $j$th row and column, $W_{11}$ the rest
            solve $W_{11}\beta - s_{12} + \lambda Sign(\beta) = 0$ using the cyclical
            ...coordinate-descent algorithm for the modified lasso
        end for
    until convergence
    for  $j = 1, 2, \ldots, p$ do
        solve $\hat{\theta}_{22} \leftarrow \frac{1}{s_{22} - w_{12}^T \hat{\beta}}$
        solve $\hat{\theta}_{12} \leftarrow -\hat{\beta} \cdot \hat{\theta}_{22}$
    end for
end procedure
procedure COORDINATEDESCENT:( $V \leftarrow W_{11}$ )
    repeat $j = 1, 2, \ldots, p - 1$
        $\hat{\beta}_j \leftarrow S(s_{12j} - \sum_{k \neq j} V_{kj}\hat{\beta}_j, \lambda)/V_{jj}$
    until convergence
end procedure                        $\#S(x, t) = sign(x)(|x| - t)_+$
```

Graphical Lasso Properties

- Computational speed
  - The graphical lasso algorithm is extremely fast
  - can solve a moderately sparse problem with 1000 nodes in less than a minute.
  - It can be modified to have edge–specific penalty parameters λ_{jk}
  - setting $\lambda_{jk} = \infty$ will force $\hat{\theta}_{jk}$ to be zero
  - graphical lasso subsumes the parameter learning algorithm.
- Missing data
  - some missing observations may be imputed by EM algorithm from the model
  - latent – fully unobserved variables – do not bring more power in Gaussian graphical model
  - latent variables are very important in discrete distributions (later).

Example



$\lambda = 36$

$\lambda = 27$

$\lambda = 7$

$\lambda = 0$

Gaussian Distribution Reparametrization (skipped)

- For a Gaussian Distribution $\phi(\mathbf{x}) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)}$
- we define
  - concentration matrix $K = \Sigma^{-1}$
  - $h = K\mu$
  - $a = -\frac{p}{2}\log(2\pi) + \frac{1}{2}log(|K|) - \frac{1}{2}\mu^T K\mu.$
- We can rewrite the join probability density to

$$\begin{aligned}
\phi(\mathbf{x}) &= (2\pi)^{-\frac{p}{2}}|K|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)K(\mathbf{x}-\mu)\right\} \\
&= (2\pi)^{-\frac{p}{2}}|K|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\mu^T K\mu + h^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T K\mathbf{x}\right\} \\
&= \exp\left\{a + h^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T K\mathbf{x}\right\} \\
&= \exp\left\{a + \Sigma_u h_u \mathbf{x}_u - \frac{1}{2}\Sigma_{u,v} K_{u,v}\mathbf{x}_u\mathbf{x}_v\right\}.
\end{aligned}$$

Gaussian Distribution Decomposition (skipped)

Lemma

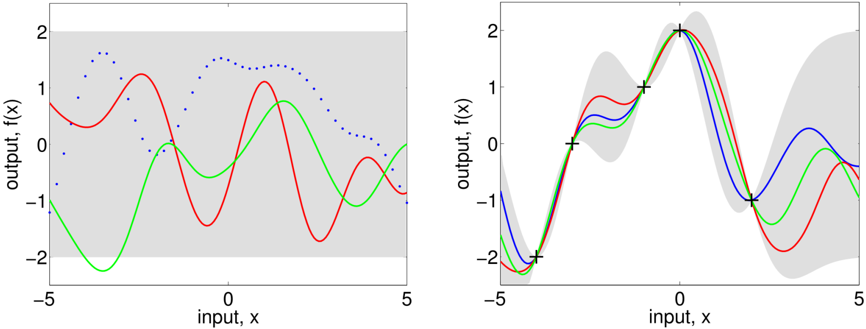
If the concentration matrix of a multivariate Gaussian distribution fulfills condition of a graph model then the distribution can be written as a product of distributions on cliques of the graph.

- $\phi(x) = \exp\left\{a + \Sigma_{u\in U} h_u \mathbf{x}_u - \frac{1}{2}\Sigma_{u,v} K_{u,v}\mathbf{x}_u\mathbf{x}_v\right\}$
- Let us have two sets of vertices A, B separated by the set C. Then $\forall u \in A, v \in B\ k_{uv} = 0$.
- We split the summation in the formula: $\phi(x) =$ $$\exp\left\{\begin{array}{c} a + \Sigma_{u\in A\cup C} h_u\mathbf{x}_u + \Sigma_{v\in B\cup C} h_v\mathbf{x}_v - \Sigma_{v\in C} h_v\mathbf{x}_v \\ -\frac{1}{2}(\Sigma_{u,v\in A\cup C} K_{u,v}\mathbf{x}_u\mathbf{x}_v + \Sigma_{u,v\in B\cup C} K_{u,v}\mathbf{x}_u\mathbf{x}_v - \Sigma_{u,v\in C} K_{u,v}\mathbf{x}_u\mathbf{x}_v) \end{array}\right\}$$
- therefore $\phi(x) = g(A, C)h(C, B)$.

$$\begin{array}{c|ccc|c|}
 & A & C & B & \\
A & K_{AA} & K_{AC} & & \\
C & K_{AC} & K_{CC} & K_{CB} & \\
B & & K_{BC} & K_{BB} &
\end{array}$$

Gaussian Processes

- An infinite (continuous) number of Gaussian variables
- to any value x a new variable $N(\mu = f(x), \Sigma_{x|rest})$
- we have only a finite number of observations which means a finite number of variables
  - we can marginalize unobserved variables out (the integral is 1, we multiply by 1, we just remove),
- we can predict at any x, continuously.



(a), prior

(b), posterior

Gaussian Processes

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006

Definition (Gaussian Process)

A Gaussian process is a set of random variables where any finite subset follows multivariate Gaussian distribution.

We define the mean $m(x)$ and the symmetric positive semidefinite covariance function $k(x, x^{\prime})$:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}^{\prime}) &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}^{\prime}) - m(\mathbf{x}^{\prime}))] \end{aligned}$$

a Gaussian process is

$$f(\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}^{\prime})).$$

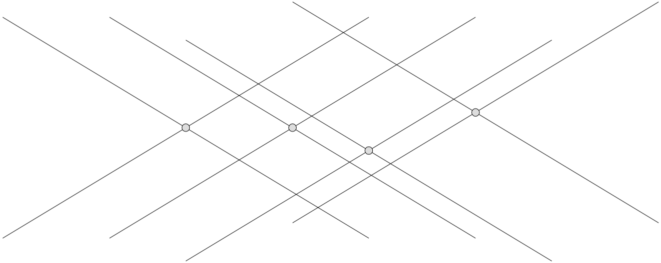
We assume $m(\mathbf{x}) = \mathbf{0}$ it simplifies the formulas.

Brownian Motion (Wiener Process)

https://www.coursera.org/lecture/stochasticprocesses/week-4-6-two-definitions-of-a-brownian-motion-THRqL

Definition (Brownian motion 1)

- $B_0 = 0$ for sure
- stationary and independent increments
- $B_s - B_t \sim N(0, s-t)$



Definition (Brownian motion 1)

- $B_0 = 0$ almost surely
- B_t stationary and independent increments
- $B_s - B_t \sim N(0, s-t)$

Definition (Brownian motion 2)

Gaussian process with

- $m = 0$ and
- $k(x, y) = min(x, y)$.

Positive semidefinite:

- $min(t, s) = \int_0^\infty f_t(x) f_s(x) dx$
- $f_t(x) f_s(x) = 1$ iff $x \in [0, t] \& x \in [0, s]$

Lemma (2⇒1)

- $K(0, 0) = min(0, 0) = 0$
- *The process has variance 0 at $t = 0$ and $m(0) = 0$.*
- *covariance is linear in both arguments, $s \geq t$*

$$\begin{aligned} cov(B_s - B_t, B_s - B_t) &= cov(B_s, B_s) - cov(B_t, B_s) - cov(B_s, B_t) + cov(B_t, \\ &= s - 2t + t = s - t \end{aligned}$$

- *increments, $s \geq t \geq b \geq a$* *# independence skipped, from Gaussian vectors*

$$\begin{aligned} cov(B_b - B_a, B_s - B_t) &= cov(B_b, B_s) - cov(B_a, B_s) - cov(B_b, B_t) + cov(B_a, B_t) \\ &= b - a - b + a = 0. \end{aligned}$$

Linear Regression as a Gaussian Process

- Consider $\phi(x)$ set of features of x (or simply the identity $\phi(x) = x$).
- Assume the function f is a linear combination of the features with weights $\mathbf{w}$.
  - Assume $m(\mathbf{x}) = \mathbf{0}$ to simplify the formula.

Linear function of features (may be omitted)

$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$ with prior $\mathbf{w} \sim N(\mathbf{0}, \Sigma_p)$.

- mean $\mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^T \mathbb{E}[\mathbf{w}] = 0$
- covariance $\mathbb{E}[f(\mathbf{x}) f(\mathbf{x}^{|})] = \phi(\mathbf{x})^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi(\mathbf{x}^{|}) = \phi(\mathbf{x})^T \Sigma_p \phi(\mathbf{x}^{|})$
  - Note that the covariance of outputs $f(\mathbf{x})$ is a function of inputs.

Corresponds to the Gaussian process with the covariance

$K(X, X^{|}) = \phi(X)^T \Sigma_p \phi(X^{|})$.

- for any number of data samples X.
- For low number of samples the covariance matrix will be singular.

Now we leave the linearity assumption and change the covariance function.

Squared exponential covariance function

Definition (Squared exponential covariance function)

Squared exponential (SE) covariance function

$$cov(f(\mathbf{x}_p), f(\mathbf{x}_q)) = k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{|\mathbf{x}_p - \mathbf{x}_q|^2}{\ell^2}\right).$$

with hyperparameters

- ℓ^2 lenghtscale,
- σ_f^2 signal variance.

- Again, the covariance on y is defined by the covariance on the inputs $\mathbf{x}$.
- the covariance defines also the distribution on functions f:

$$\mathbf{f}_* \sim N(\mathbf{0}, K(X_*, X_*)).$$

Prediction

- noisy-free prediction $y = f(\mathbf{x})$

$$cov(y_p, y_q) \quad = \quad k(\mathbf{x}_p, \mathbf{x}_q)$$

- from noisy observations $y = f(\mathbf{x}) + \epsilon$, $e \sim N(0, \sigma_n^2)$

$$\begin{aligned} cov(y_p, y_q) &= k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_n^2 \delta_{pq} \\ cov(\mathbf{y}) &= K(X, X) + \sigma_n^2 I \end{aligned}$$

- We observe $\mathbf{y}$ and we want to predict $\mathbf{f}_*$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix})$$

- Predictive distribution

$$\begin{aligned} \mathbf{f}_* | X, \mathbf{y}, X_* &\sim N(\bar{\mathbf{f}}_*, cov(\bar{\mathbf{f}}_*)) \\ \bar{\mathbf{f}}_* &\triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} \mathbf{y} \\ cov(\mathbf{f}_*) &= K(X_*, X_*) - K(X_*, X)[K(X_*, X) + \sigma_n^2 I]^{-1} K(X, X_*) \end{aligned}$$

Predictive distribution

$$\begin{aligned}
\mathbf{f}_*|X,\mathbf{y},X_* &\sim N(\bar{\mathbf{f}}_*, cov(\bar{\mathbf{f}}_*)) \\
\bar{\mathbf{f}}_* &\triangleq \mathbb{E}[\mathbf{f}_*|X,\mathbf{y},X_*] = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y} \\
cov(\mathbf{f}_*) &= K(X_*,X_*) - K(X_*,X)[K(X_*,X)+\sigma_n^2 I]^{-1}K(X,X_*)
\end{aligned}$$

- We denote $K = K(X,X)$, $K_*(X,X_*)$, for a single variable $\mathbf{k}_* = \mathbf{k}(\mathbf{x},\mathbf{x}_*)$. Then for $\mathbf{x}_*$:

$$\begin{aligned}
\bar{f}_* &\triangleq \mathbf{k}_*^T[K+\sigma_n^2 I]^{-1}\mathbf{y} \\
\mathbb{V}(f_*) &= \mathbf{k}(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^T[K+\sigma_n^2 I]^{-1}\mathbf{k}_*
\end{aligned}$$

Prediction

- is a linear function of observations
  - for $\alpha \Leftarrow (K+\sigma_n^2 I)^{-1}\mathbf{y}$
  - we predict $\bar{f}(\mathbf{x}_*) \Leftarrow \sum_{i=1}^n \alpha_i k(\mathbf{x}_i,\mathbf{x}_*)$

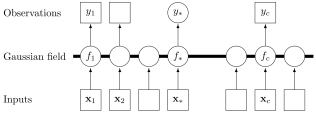


Figure 2.3: Graphical model (chain graph) for a GP for regression. Squares represent observed variables and circles represent unknowns. The thick horizontal bar represents a set of fully connected nodes. Note that an observation y_i is conditionally independent of all other nodes given the corresponding latent variable, f_i. Because of the marginalization property of GPs addition of further inputs, $\mathbf{x}$, latent variables, f, and *unobserved* targets, y_*, does not change the distribution of any other variables.

Marginal likelihood (may be omitted)

- 'In sample' prediction **f** follows:

$$\mathbf{f} \sim N(\mathbf{0}, K(X, X))$$

- Noisy-free observations $\mathbf{y} = \mathbf{f}$:

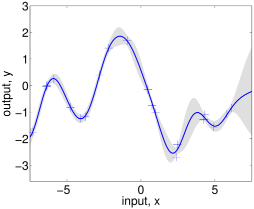
$$\log p(\mathbf{y}|X) = \log p(\mathbf{f}|X) = -\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

- For noisy observations $\mathbf{y}|\mathbf{f} \sim N(\mathbf{f}, \sigma_n^2 I)$, $\mathbf{y} \sim N(0, K + \sigma_n^2 I)$

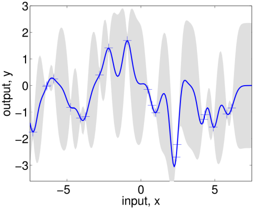
$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi.$$

Hyperparameters ℓ, σ_f

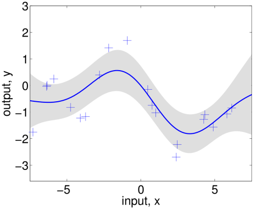
- a. Data are generated by a GP with SE $(\ell, \sigma_f, \sigma_n) = (1, 1, 0.1)$ (lenghtscale, signal variance, noise variance)
- b. 95% confidence intervals for (0.3,1.08,0.00005)
- c. 95% confidence intervals for (3.0,1.16,0.89).



(a), $\ell = 1$

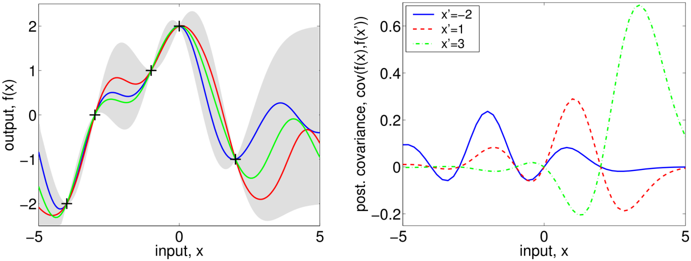


(b), $\ell = 0.3$



(c), $\ell = 3$

Conditional Covariance (may be omitted)



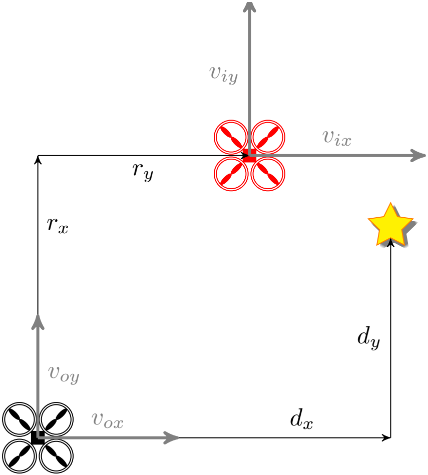
(a), posterior

(b), posterior covariance

Figure 2.4: Panel (a) is identical to Figure 2.2(b) showing three random functions drawn from the posterior. Panel (b) shows the posterior *co*-variance between $f(\mathbf{x})$ and $f(\mathbf{x}')$ for the same data for three different values of $\mathbf{x}'$. Note, that the covariance at close points is high, falling to zero at the training points (where there is no variance, since it is a noise-free process), then becomes negative, etc. This happens because if the smooth function happens to be less than the mean on one side of the data point, it tends to exceed the mean on the other side, causing a reversal of the sign of the covariance at the data points. Note for contrast that the *prior* covariance is simply of Gaussian shape and never negative.

POMDP Aircraft Collision Avoidance

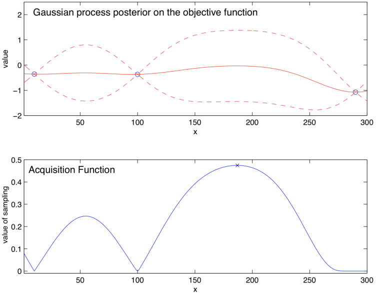
- multirotor aircraft (drones) and helicopters actions
    - turns
    - vertical meneuvers
    - horizontal plane accelerations
- QMDP with hyper–parameters
    - K_s, K_T, R_{min} weights was learned, k weights was $= 1$.
    - The performance for a specific hyper-parameters is costly to evaluate.



$$R(s,a) = max\left[R_{min}, -(k_{ax}|a_x| + k_{ay}|a_y|) - K_s\frac{1}{k_{rx}r_x^2 + k_{ry}r_y^2} - K_T(k_{dx}d_x^2 + k_{dy}d_y^2)\right]$$

Bayesian Optimization

- Simulate process and evaluate collision probability $F(K_s, K_T, R_{min})$ at a minimal number of points
- learn a Gaussian model
- find the most promising values to evaluate next
    - we minimize $y = F(R_P^*)$ and search the maximal probability of improvement
    - 'the chance to improve' is expressed by the **Expected improvement** (EI)



Peter I. Frazier: A Tutorial on Bayesian Optimization, rXiv:1807.02811v1 [stat.ML] 8 Jul 2018

Bayesian Optimization

- The **Expected improvement** (EI) is defined to be:

$$\begin{aligned} EI(x) &= \mathbb{E}[(min(Y(X)) - Y(x))^+ | Y(X) = \mathbf{y}] \\ &= \mathbb{E}[(min(\mathbf{y}) - Y(x))^+ | Y(X) = \mathbf{y}] \end{aligned}$$

- this can be solved analytically (Φ cdf, ϕ pdf Gaussian distribution):

$$EI(x) = (min(\mathbf{y}) - \mu(x))\Phi\left(\frac{min(\mathbf{y}) - \mu(x)}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{min(\mathbf{y}) - \mu(x)}{\sigma(x)}\right)$$

- to maximize $\mathbf{y}$:

$$EI(x) = \Delta(x)^+\sigma(x)\phi\left(\frac{\Delta(x)}{\sigma(x)}\right) - |\Delta(x)|\Phi\left(\frac{\Delta(x)}{\sigma(x)}\right)$$

$(\Delta(x) = \mu(x) - max(\mathbf{y})\ )$

Covariance Functions

- radial basis functions RBF $\varphi(r)$, for $r = |\mathbf{x} - \mathbf{x}^{|}|$
- Modified Bessel functions (for α not integer, the limit otherwise)
  - $I_\alpha(x) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+\alpha+1)} \left(\frac{x}{2}\right)^{2m+\alpha}$
  - $K_\alpha(x) = \frac{\pi}{2} \frac{I_{-\alpha}(x) - I_\alpha(x)}{sin\alpha\pi}$

covariance function	expression	Stacionary	inf. rank
constant	σ_0^2	y	
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x_d^{|}$		
polynomial	$(\mathbf{x} \cdot \mathbf{x}^{|} + \sigma_0^2)^p$	y	
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$	y	y
Matérn, $\nu = k + \frac{1}{2}$	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$	y	y
exponential	$\exp(-\frac{r}{\ell})$	y	y
γ-exponential	$\exp(-(\frac{r}{\ell})^{\gamma})$	y	y
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$	y	y
neural network	$\sin^{-1}\left(\frac{2\widetilde{\mathbf{x}}^T \Sigma \widetilde{\mathbf{x}}^{|}}{\sqrt{(1+2\widetilde{\mathbf{x}}^T \Sigma \widetilde{\mathbf{x}})(1+2\widetilde{\mathbf{x}^{|}}^T \Sigma \widetilde{\mathbf{x}^{|}})}}\right)$	y	y

Table of Contens

1. Overview of Supervised Learning
2. Undirected (Pairwise Continuous) Graphical Models
3. Gaussian Processes, Kernel Methods
4. Basis Expansion and regularization, Splines
5. Linear methods for classification
6. Model Assessment and Selection
7. Bayesian learning, EM algorithm
8. Decision trees, MARS, PRIM
9. Ensamble Methods
10. Association Rules, Apriori
11. Clustering
12. Summary