Spatial Statistics with Image Analysis
Lecture 10

Johan Lindström

December 1, 2017

Object classification

- We assume that there is a set of $K$ object categories.
- We take the latent variables $z_i = z(s_i)$ to be the (random) category for pixel $y_i$ at location $s_i$.
- The prior probabilities $\pi_k$ describe the relative abundance of each object class.
- For each object class, there is a data model/distribution, $p(y_i | z_i = k)$.
Bayesian hierarchical modelling (BHM)

A hierarchical model is constructed by systematically considering components/features of the data, and how/why these features arise.

Bayesian hierarchical modelling

A Bayesian hierarchical model typically consists of (at least)

Data model, \( p(y|z, \theta) \): Describing how observations arise given the latent variables \( z \) and parameters \( \theta \).

Latent model, \( p(z|\theta) \): Describing how the latent variables (reality) behaves, given parameters.

Parameters, \( p(\theta) \): Describing our, sometimes vague, prior knowledge of the parameters.

Object classification as a BHM

Treating the classification problem as a BHM we have

Data model, \( p(y|z, \theta) \): Distribution of a pixel given its class belonging \( z \) and parameters \( \{\mu_k, \Sigma_k\}_{k=1}^K \).

\[ [y_i | z_i = k] \in \mathcal{N}(\mu_k, \Sigma_k) . \]

Latent model, \( p(z|\theta) \): Describing how common each class is given parameters \( \{\pi_k\}_{k=1}^K \).

\[ p(z_i = k) = \pi_k, \quad \forall i . \]

Parameters, \( p(\theta) \): Describing our vague prior knowledge of the parameters.

\[ \theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K, \quad p(\theta) \propto 1. \]

Object classification using a Gaussian mixture model

The resulting model is also called a Gaussian mixture model since the density for \( y_i \) is

\[ p(y_i|\theta) = \sum_{k=1}^K \pi_k \left( \frac{1}{2\pi} \right)^{d/2} |\Sigma_k|^{1/2} \exp \left( -\frac{1}{2} \frac{1}{\Sigma_k}(y_i - \mu_k)^\top \Sigma_k^{-1}(y_i - \mu_k) \right) . \]

Direct maximum-likelihood estimation is hard since

\[ \log p(y|\theta) = \sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k \cdot p_G(y_i|\mu_k, \Sigma_k) \right) . \]
**Parameter estimation in a GMM**

The basic idea of the parameter estimation is to reduce the problem into two parts:

1. Determine the class belongings \( z \) (given parameters)
   \[
   p(z_i = k | y_i, \theta) = \frac{p(y_i | z_i = k, \theta) \pi_k}{\sum_l p(y_i | z_i = l, \theta) \pi_l}
   \]

2. Estimating the parameters \( \theta \) (given known class belongings)
   \[
   n_k = \sum_{i: z_i = k} 1 \\
   \hat{\pi}_k = \frac{n_k}{\sum_k n_k} \\
   \hat{\mu}_k = \frac{1}{n_k} \sum_{i: z_i = k} y_i \\
   \hat{\Sigma}_k = \frac{1}{n_k - d} \sum_{i: z_i = k} (y_i - \hat{\mu}_k)^\top (y_i - \hat{\mu}_k)
   \]

**Markov Chain Monte Carlo**

- Basic idea: To sample from a density \( p(\theta | y) \) we construct a Markov chain with \( p(\theta | y) \) as it’s stationary distribution.
- The samples will not be independent.
- MCMC is one of the most common method for inference in more complex spatial models.
- Dates back to the 1950's with two key papers being:
  - Equations of State Calculations by Fast Computing Machines (1953)

**Gibbs sampling**

- Divide into blocks and sample from each block conditional on the rest.
- The Gibbs sampler is a special case of MCMC.
- When sampling complex models we can often expand the posterior to include the unknown latent fields
  \[
  p(\theta, z | y)
  \]
- This has two advantages
  1. It provides a direct reconstruction of the latent field.
  2. (Often) Easier to sample from \( p(\theta | z, y) \) and \( p(z | \theta, y) \) than \( p(\theta | y) \).
- Given a joint sample from \( [\theta, z | y] \) a sample from \( [\theta | y] \) is simply obtained by ignoring the \( z \) values.
**Gibbs sampling**

**Algorithm:**
1. Choose a starting value \( \theta^{(0)} \).
2. Repeat for \( i = 1, \ldots, N \):
   - i.1 Draw \( \theta^{(i)}_1 \) from \( p(\theta_1 | \theta^{(i-1)}_2, \ldots, \theta^{(i-1)}_m, y) \).
   - i.2 Draw \( \theta^{(i)}_2 \) from \( p(\theta_2 | \theta^{(i)}_1, \theta^{(i-1)}_3, \ldots, \theta^{(i-1)}_m, y) \).
   - i.3 Draw \( \theta^{(i)}_3 \) from \( p(\theta_3 | \theta^{(i)}_1, \theta^{(i)}_2, \theta^{(i-1)}_4, \ldots, \theta^{(i-1)}_m, y) \).
   - \vdots
   - i.m Draw \( \theta^{(i)}_m \) from \( p(\theta_m | \theta^{(i)}_1, \theta^{(i)}_2, \ldots, \theta^{(i-1)}_{m-1}, y) \).
3. \( \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)} \), is now a sequence of dependent draws approximately from \( p(\theta | y) \).

**Hybrid methods**

- Divide the parameter vector into blocks and use different updating strategies for each block.
- Allows us to update several parameters in one **Gibbs-block** if we can find the joint conditional distribution.
- Allows us to use Gibbs sampling for part of the updates and other MCMC strategies such as **Metropolis Hastings** for the reminder.

**Gamma & inverse-Gamma**

\[
\begin{align*}
\Gamma(\alpha, \beta) & \quad \text{if } \Gamma(\alpha, \beta) \ni x, \\
\Gamma(x, \beta) & \quad \text{if } \Gamma(x, \beta) \ni 1/x.
\end{align*}
\]

- **f(x)**: \( \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta/x}, \ x \geq 0 \)
- **E(x)**: \( \frac{\beta}{\alpha} \)
- **x_{\text{mode}}**: \( \frac{\alpha-1}{\beta}, \ \alpha > 1 \)
- **V(x)**: \( \frac{\beta^2}{\alpha-1} \)

If \( x \in \Gamma(\alpha, \beta) \) then \( 1/x \in \Gamma(\alpha, \beta) \).

The Gamma distribution can be used to model precision, inverse-Gamma to model variances.
**Gamma — Different parameterizations**

There are two common ways of expressing the Gamma (and inverse-Gamma) distribution, scale or rate.

<table>
<thead>
<tr>
<th></th>
<th>Scale (MATLAB)</th>
<th>Rate (Here &amp; HA3)</th>
</tr>
</thead>
</table>
| $f(x)$   | $\frac{1}{
\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$, $x \geq 0$ | $\frac{\beta^\alpha}{
\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$, $x \geq 0$ |
| $E(x)$   | $k\theta$     | $\frac{\alpha}{\beta}$ |
| $X_{mode}$ | $(k-1)\theta$, $k > 1$ | $\frac{\alpha-1}{\beta}$, $\alpha > 1$ |
| $V(x)$   | $k\theta^2$   | $\frac{\alpha}{\beta^2}$ |

**Wishart**

The Wishart distributions can be seen as multi-variate version of Gamma and inverse-Gamma.

In the following $X$ and $V$ are $d \times d$ positive definite, symmetric matrices. If $X \in W(V, n)$ then

$$f(X) = \frac{|X|^{n-d-1}}{2^{nd/2} |V|^{n/2} \Gamma_d(\frac{n}{2})} e^{-\text{tr}(V^{-1}X)/2}$$

$$E(X) = nV$$

$$X_{mode} = (n-d-1)V$$

The Wishart distribution can be used to model precision matrices.

**inverse-Wishart**

If $X \in IW(V, n)$ then

$$f(X) = \frac{|V|^{n/2}}{2^{nd/2} \Gamma_d(\frac{n}{2})} |X|^{-\frac{n+d+1}{2}} e^{-\text{tr}(V^{-1}X)/2}$$

$$E(X) = \frac{V}{n-d-1}$$. $n > d + 1$

$$X_{mode} = \frac{V}{n + d + 1}$$

If $X \in W(V, n)$ then $X^{-1} \in IW(V^{-1}, n)$.

The inverse-Wishart distribution can be used to model variance matrices.

Available in matlab as wishrnd and iwishrnd.

$$\Gamma_d(a) = \frac{\pi^{d(d-1)/4}}{\prod_{i=1}^{d} \Gamma(a + (i-j)/2)}.$$
Dirichlet distribution

**Dirichlet distributions** can be used to model probability vectors.

In the following \( x \) is a \( 1 \times K \) vector of probabilities with \( 0 \leq x_i \leq 1 \) and \( \sum_i x_i = 1 \).

If \( x \in \text{Dir}(\alpha) \), with \( \alpha_i > 0 \), then

\[
f(x) = \frac{\Gamma \left( \sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma (\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \quad E(x_i) = \frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k}
\]

Multivariate case of a \( \beta(a, b) \) distribution

\[
f(x) = \frac{\Gamma (a + b)}{\Gamma (a) \Gamma (b)} x_a^{a-1} (1 - x)^{b-1} \quad E(x) = \frac{a}{a + b}
\]

Parameter estimation in a GMM

Gibbs-sampling for the GMM can now be performed by alternating

1. Sample from the class belongings \( z \) (given parameters)

\[
p(z_i = k|y_i, \theta) = \frac{p(y_i|z_i = k) \pi_k}{\sum_l p(y_i|z_i = l) \pi_l}
\]

2. Sample from the parameters \( \theta \) (given known class belongings).

Given the posterior probabilities \( p(z_i = k|y_i, \theta) \), sampling from \( z_i \) is only a case of sampling from a discrete distribution \( 1, \ldots, K \) with some probability of falling in each class. (Equivalent to rolling a weighted die).

Parameter estimation in a GMM

The second step is slightly harder. First we see that given pixel belongings \( z \) we can divide the pixels into groups

\( y^{(k)} = \{y_i : z_i = k\} \)

with \( n_k = \sum_l I(z_l = k) \) pixels in each group.

The posterior for \( \theta | z, y \) is now

\[
p(\theta | z, y) = \frac{p(y, z, \theta) p(z|\theta) p(\theta)}{p(z, y)} \propto p(y|z, \theta) p(z|\theta)
\]

\[
= \prod_{i=1}^{N} p(y_i|z_i, \theta) p(z_i|\theta) = \prod_{i=1}^{N} \pi_{z_i} p_G \left( y_i | \mu_{z_i}, \Sigma_{z_i} \right)
\]
Parameter estimation in a GMM

The posterior for all parameters is

$$p(\theta | z, y) \propto \prod_{i=1}^{N} \pi_i p_G(y_i | \mu_{z_i}, \Sigma_{z_i})$$

where $\mu_{z_i}$ is the mean value of the class indicated by $z_i$ (i.e. the sampled class belonging of $y_i$).

Dividing the problem into smaller sub-groups gives, for $\pi$,

$$p(\pi | \mu, \Sigma, z, y) \propto \prod_{i=1}^{N} \pi_i p_G(y_i | \mu_{z_i}, \Sigma_{z_i}) \propto \prod_{i, z_i = k} p_G(y_i | \mu_k, \Sigma_k)$$

$$\pi | z \in \text{Dir}(n_k + 1)$$

This is equivalent to sampling each $\mu_k, \Sigma_k$ independently, given a set of observations $y^{(k)}$ from $N(\mu_k, \Sigma_k)$.

The resulting Gibbs algorithm is:

1. Sample from the class belongings $z$

   $$p(z_i = k | y_i, \theta) = \frac{p(y_i | z_i = k) \pi_k}{\sum_l p(y_i | z_i = l) \pi_l}$$

2. Sample the class proportions $\pi_k$

   $$\pi | z \in \text{Dir}(n_k + 1)$$

3. Sample the class parameters $\mu_k, \Sigma_k$ from

   $$p(\mu_k, \Sigma_k | z(k)) \propto \prod_{i, z_i = k} p_G(y_i | \mu_k, \Sigma_k)$$
The class parameters — $\mathbf{N}(\mu, \sigma^2)$

Assume we have $n$ independent observations, $y_i$ from a Gaussian distribution with flat priors for $\mu$ and $\sigma^2$.

$$y_i | \mu, \sigma^2 \sim \mathbf{N}(\mu, \sigma^2) \quad \text{p}(\mu, \sigma^2) \propto 1$$

The posterior for the parameters is

$$p(\mu, \sigma^2 | \mathbf{y}) = \frac{p(\mathbf{y} | \mu, \sigma^2) p(\mu, \sigma^2)}{p(\mathbf{y})} \propto \prod_{i=1}^{n} p(y_i | \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \mu)^2}{2\sigma^2} \right)$$

$$\propto (\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right)$$

Studying the posterior

$$p(\mu, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp \left( -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right)$$

we see that

$$[\mu | \sigma^2, \mathbf{y}] \in \mathbf{N} \left( \frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sigma^2}{n} \right)$$

and

$$[\sigma^2 | \mu, \mathbf{y}] \in \mathbf{IG} \left( \frac{n}{2} - 1, \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} \right)$$

Class parameters — Results

Given $N = 50$ independent observations from a $\mathbf{N}(5, 2)$-distribution we estimate $\mu$ and $\sigma^2$ using Gibbs-sampling.
Joint sample from $\mu, \sigma^2 | y$

On the previous slides we have sampled the posterior class parameters by a two-step Gibbs sample:

$$\begin{align*}
\left[ \mu \right| \sigma^2, y &\in \mathcal{N}\left( \frac{\sum_{i=1}^{n} y_i}{n}, \frac{\sigma^2}{n} \right) \\
\left[ \sigma^2 \right| \mu, y &\in \mathcal{I}\Gamma \left( \frac{n}{2} - 1, \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} \right)
\end{align*}$$

A better option would be to sample $\mu$ and $\sigma^2$ jointly (this improves the convergence speed in the Gibbs sampling). To sample $\mu$ and $\sigma^2$ jointly we rewrite the posterior as

$$p(\mu, \sigma^2 | y) = p(\mu | \sigma^2, y) p(\sigma^2 | y) .$$

and the joint sample is compute by first sampling $[\sigma^2 | y]$ followed by $[\mu | \sigma^2, y]$.

Completing the squares in the exponent wrt to $\mu$ gives

$$\begin{align*}
-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2 &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{n}{\sigma^2} \sum_{i=1}^{n} y_i - n \mu^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{n}{\sigma^2} \sum_{i=1}^{n} y_i - n \frac{\mu^2}{2}\sigma^2
\end{align*}$$

where $\bar{y} = (\sum_{i=1}^{n} y_i)/n$. We recognise the second part as the exponential of a Gaussian distribution for $\mu$. 

Finding the posteriors $[\sigma^2 | y]$ followed by $[\mu | \sigma^2, y]$, is equivalent to factoring the density $[\mu, \sigma^2 | y]$ into one component containing $\mu$ and $\sigma^2$ and one component containing only $\sigma^2$.

$$p(\mu, \sigma^2 | y) \propto \frac{1}{(\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right)$$

Here the exponent is

$$\begin{align*}
-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2 &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{n}{\sigma^2} \sum_{i=1}^{n} y_i - n \mu^2 \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{n}{\sigma^2} \bar{y} + \frac{n}{n} \sum_{i=1}^{n} y_i - n \frac{\mu^2}{2}\sigma^2
\end{align*}$$
Joint sampling

For the first part (i.e. parts not containing $\mu$) we write

$$- \frac{1}{2 \sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{n}{2 \sigma^2} y^2 = - \frac{1}{2 \sigma^2} \left( \sum_{i=1}^{n} y_i^2 - ny^2 \right)$$

$$= - \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (y_i - y)^2$$

Since

$$\sum_{i=1}^{n} (y_i - y)^2 = \sum_{i=1}^{n} (y_i^2 - 2yy_i + y^2)$$

$$= \sum_{i=1}^{n} y_i^2 - 2y \sum_{i=1}^{n} y_i + ny^2 = \sum_{i=1}^{n} y_i^2 - ny^2$$

We can now write the exponent as

$$- \frac{1}{2 \sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = - \frac{S_{yy}}{2 \sigma^2} - \frac{n}{2 \sigma^2} (\mu - \bar{y})^2$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The density becomes

$$p(\mu, \sigma^2 | y) \propto \frac{1}{(\sigma^2)^{n/2}} \exp \left( - \frac{S_{yy}}{2 \sigma^2} - \frac{n}{2 \sigma^2} (\mu - \bar{y})^2 \right)$$

$$= \frac{1}{(\sigma^2)^{(n-1)/2}} \exp \left( - \frac{S_{yy}}{2 \sigma^2} \right) \cdot \frac{1}{\Gamma\left(\frac{n+1}{2}\right)} \exp \left( - \frac{n}{2 \sigma^2} (\mu - \bar{y})^2 \right)$$

$$\cdot \Gamma\left(\frac{n+1}{2} - 1\right) \cdot \frac{1}{\sigma^2}$$

$$N\left(\bar{y}, \frac{1}{\sigma^2}\right)$$

The joint sampling is now accomplished by first sampling $\sigma^2 | y$ (from an inverse-Gamma) followed by sampling $\mu | \sigma^2, y$ (from a Normal).
Classification MCMC Distributions Estimation MRFs

Class parameters — $N(\mu, \Sigma)$

For a multivariate normal, $y_i \in N(\mu, \Sigma)$ we have

$$p(\mu, \Sigma | y) \propto \frac{1}{|\Sigma|^{n/2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) \right)$$

The posterior for $\mu$ is

$$[\mu | \Sigma, y] \in N \left( \frac{\sum_{i=1}^{n} y_i}{n}, \Sigma \right).$$

The posterior for $\Sigma$ requires some effort.

Properties of trace

For the trace,

$$\text{tr}(A) = \sum_i A_{ii},$$

of a matrix we have the following properties

Cyclic permutations:

$$\text{tr} (ABC) = \text{tr} (BCA) = \text{tr} (CAB)$$

Summation:

$$\text{tr} (A + B) = \text{tr} (A) + \text{tr} (B)$$

$$\text{tr} (A(B + C)) = \text{tr} (AB) + \text{tr} (AC)$$

Class parameters — $N(\mu, \Sigma)$

Using the computational rules for trace we have

$$\sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) = \sum_{i=1}^{n} \text{tr} \left( (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) \right)$$

$$= \sum_{i=1}^{n} \text{tr} \left( \Sigma^{-1} (y_i - \mu) (y_i - \mu)^\top \right)$$

$$= \text{tr} \left( \Sigma^{-1} \left[ \sum_{i=1}^{n} (y_i - \mu) (y_i - \mu)^\top \right] \right)$$

and the posterior for $\Sigma$ is

$$[\Sigma | \mu, y] \in \text{IW} \left( \sum_{i=1}^{n} (y_i - \mu) (y_i - \mu)^\top, n - d - 1 \right)$$
Joint sample from $\mu, \Sigma \mid y$

Just as for the univariate case replacing the two-step Gibbs sample:

$$[\mu \mid \Sigma, y] \sim N \left( \frac{\sum_{i=1}^{n} y_i}{n}, \frac{\Sigma}{n} \right)$$

$$[\Sigma \mid \mu, y] \sim \text{IW} \left( \frac{n}{\sum_{i=1}^{n} (y_i - \mu)^\top (y_i - \mu)} n - d - 1 \right).$$

with joint samples from $\mu$ and $\Sigma$ would improve convergence in the Gibbs sampling.

To sample $\mu$ and $\Sigma$ jointly we rewrite the posterior as

$$p(\mu, \Sigma \mid y) = p(\mu \mid \Sigma, y) p(\Sigma \mid y).$$

and the joint sample is compute by first sampling $[\Sigma \mid y]$ followed by $[\mu \mid \Sigma, y]$.

Finding the posteriors $[\Sigma \mid y]$ and $[\mu \mid \Sigma, y]$ is equivalent to factoring the density $[\mu, \Sigma \mid y]$ into one component containing $\mu$ and $\Sigma$ and one component containing only $\Sigma$.

$$p(\mu, \Sigma \mid y) \propto \frac{1}{\sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) \right)$$

Here the sum in the exponent is

$$\sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) = \sum_{i=1}^{n} y_i^\top \Sigma^{-1} y_i - 2\mu^\top \left( \frac{\Sigma}{n} \right)^{-1} \sum_{i=1}^{n} y_i + \mu^\top \left( \frac{\Sigma}{n} \right)^{-1} \mu.$$

Completing the squares in the exponent wrt $\mu$ gives

$$\sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) = \sum_{i=1}^{n} (y_i^\top \Sigma^{-1} y_i) - y^\top \left( \frac{\Sigma}{n} \right)^{-1} y$$

$$- 2\mu^\top \left( \frac{\Sigma}{n} \right)^{-1} y + \mu^\top \left( \frac{\Sigma}{n} \right)^{-1} \mu + y^\top \left( \frac{\Sigma}{n} \right)^{-1} \mu - y^\top \left( \frac{\Sigma}{n} \right)^{-1} \mu + (\mu - y)^\top \left( \frac{\Sigma}{n} \right)^{-1} (\mu - y)^\top.$$

where $y = \left( \sum_{i=1}^{n} y_i \right) / n$. We recognise the second part as the exponential of a Gaussian distribution for $\mu$. 

Johan Lindström - johanl@maths.lth.se
Joint sampling

The first part we rewrite as:

\[
\sum_{i=1}^{n} (y_i - \mu) \Sigma^{-1} (y_i - \mu) = \text{tr} \left( \Sigma^{-1} \sum_{i=1}^{n} (y_i y_i^\top) \right) - \text{tr} \left( \Sigma^{-1} \left[ n \bar{y} \bar{y}^\top \right] \right)
\]

\[
= \text{tr} \left( \Sigma^{-1} \sum_{i=1}^{n} (y_i y_i^\top) - n \bar{y} \bar{y}^\top \right)
\]

\[
= \text{tr} \left( \Sigma^{-1} \sum_{i=1}^{n} (y_i - \bar{y}) (y_i - \bar{y})^\top \right).
\]

Thus we can write the exponent as

\[
\sum_{i=1}^{n} (y_i - \mu)^\top \Sigma^{-1} (y_i - \mu) = \text{tr} \left( \Sigma^{-1} s_{yy} \right) + (\mu - \bar{y})^\top \left( \frac{\Sigma}{n} \right)^{-1} (\mu - \bar{y})^\top,
\]

where

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y}) (y_i - \bar{y})^\top.
\]

The density becomes

\[
p(\mu, \Sigma | y) \propto \frac{1}{|\Sigma|^n/2} \exp \left( -\frac{1}{2} \text{tr} \left( \Sigma^{-1} s_{yy} \right) \right) \cdot \exp \left( -\frac{1}{2} (\mu - \bar{y})^\top \left( \frac{\Sigma}{n} \right)^{-1} (\mu - \bar{y}) \right)
\]

\[
= \frac{1}{|\Sigma|^{n/2} \Gamma_n} \exp \left( -\frac{1}{2} \text{tr} \left( \Sigma^{-1} s_{yy} \right) \right) \cdot \frac{1}{n \Gamma_{n/2} \left( n - d - 2 \right)} \cdot \frac{1}{n(\pi, \frac{\bar{y}}{\bar{y}})}.
\]

The joint sampling is now accomplished by first sampling \( \Sigma | y \) (from an inverse-Wishart) followed by sampling \( \mu | \Sigma, y \) (from a Normal).
General Markov random fields

In the classification above we have assumed that the class belonging of each pixels is independent of the class belonging of other pixels. To allow for dependence between different pixels we need to define a field for discrete values.

Gaussian Random Field \( x \in \mathbb{N}(\mu, \Sigma) \) with elements of the covariance matrix, \( \Sigma \), defined through a covariance function.

Gaussian Markov Random Field \( x \in \mathbb{N}(\mu, Q^{-1}) \) with sparse precision matrix \( Q \).

Markov Random Field \( x \in \mathbb{N} \) such that

1. \( x \) is Markov.
2. \( x \) takes values in a suitable set \( x_i \in 1, 2, \ldots, K \).

We want models, defined for locations \( s_i \), which fulfil the Markov condition

\[ p(\{x_i|\{x_j:j \neq i\}) = p(\{x_i|\{x_j:j \in N_i\}) \]

for the neighbours \( N_i \).

Gibbs distributions

Originally from mathematical physics:

\[ p(x) = \frac{1}{Z} \exp \left( -W(x)/(kT) \right), \]

where \( W \) is the energy of a state \( x \), \( T \) is a temperature, \( k \) is some physical constant, and \( Z \) is an (unknown) normalising constant.

Cliquies

The energy function will be constructed in a special way, using the concept of cliques on a given neighbourhood system.

Clique

Any single site or any set of sites, all distinct pairs of which are neighbours is called a clique.

All points in a (non-trivial) clique are neighbours of each-other

\[ s_i, s_j \in \mathcal{C} \Rightarrow j \in N_i, \]

and for any two points that are neighbours, there exists (at least one) clique containing the points

\[ j \in N_i \Rightarrow \exists \mathcal{C}: s_i, s_j \in \mathcal{C}. \]
Gibbs distributions

- The energy function $W(x)$ is constructed as a sum of potentials over all cliques,

$$W(x) = - \sum_C V_C(x)$$

where each potential function $V_C$ only depends on the $x(s_j)$'s belonging to the clique $C$.

- For stationary fields, the potentials only depend on the type of clique, not its location.

Gibbs distributions (cont)

- Neighbourhood system $\mathcal{N}_j = \{s_j', s_j' \text{ is a neighbour of } s_j\}$.
- Cliques $C$.
- Potentials $V_C(x, \theta)$.
- Parameters $\theta$.
- Density/probability function

$$p(x|\theta) = \frac{1}{Z_{\theta}} \exp \left( \sum_C V_C(x; \theta) \right)$$

- Conditional distributions

$$p(x_i|x_j, j \in \mathcal{N}_i, \theta) = \frac{1}{Z_{i, \theta}} \exp \left( \sum_{C: s_j \in C} V_C(x; \theta) \right)$$