Bayesian Methods and Markov Random Fields

Mário A. T. Figueiredo

Department of Electrical and Computer Engineering

Instituto Superior Técnico

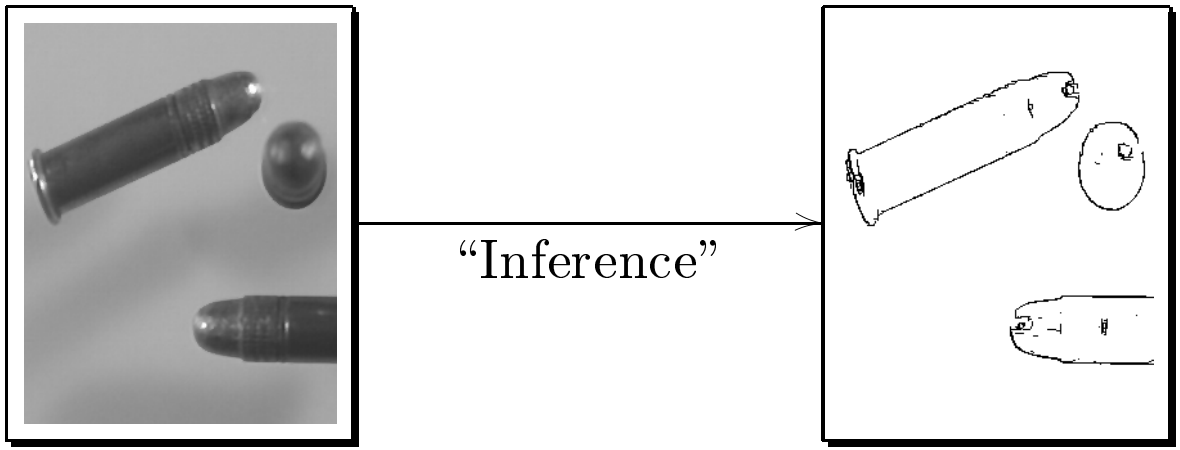
Lisboa, PORTUGAL

email: mtf@lx.it.pt

Thanks: *Anil K. Jain* and *Robert D. Nowak*, Michigan State University, USA

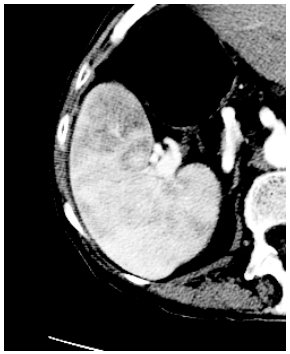
Most image analysis problems are "inference" problems:

For example, "edge detection":

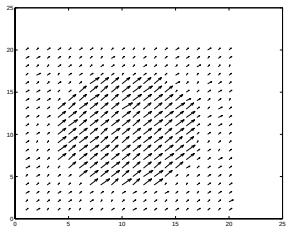


The word "image" should be understood in a wide sense. Examples:

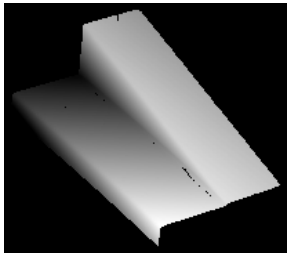
Conventional image



CT image



Flow image



Range image

Examples of "inference" problems

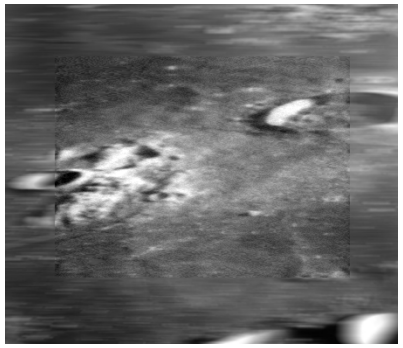
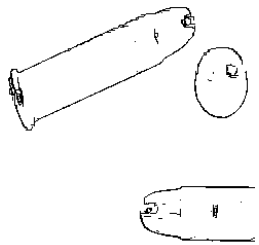
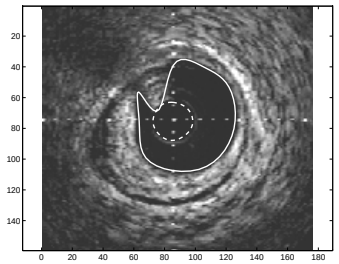


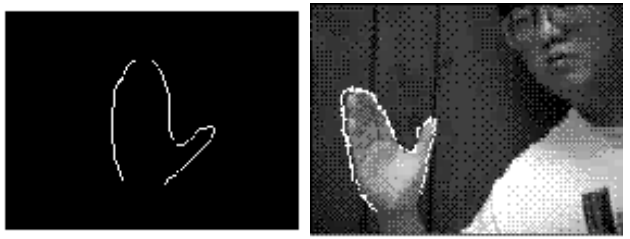
Image restoration



Edge detection



Contour estimation



Template matching

Main features of "image analysis" problems

- They are inference problems, i.e., they can be formulated as:

  "from $\mathbf{g}$, infer $\mathbf{f}$"

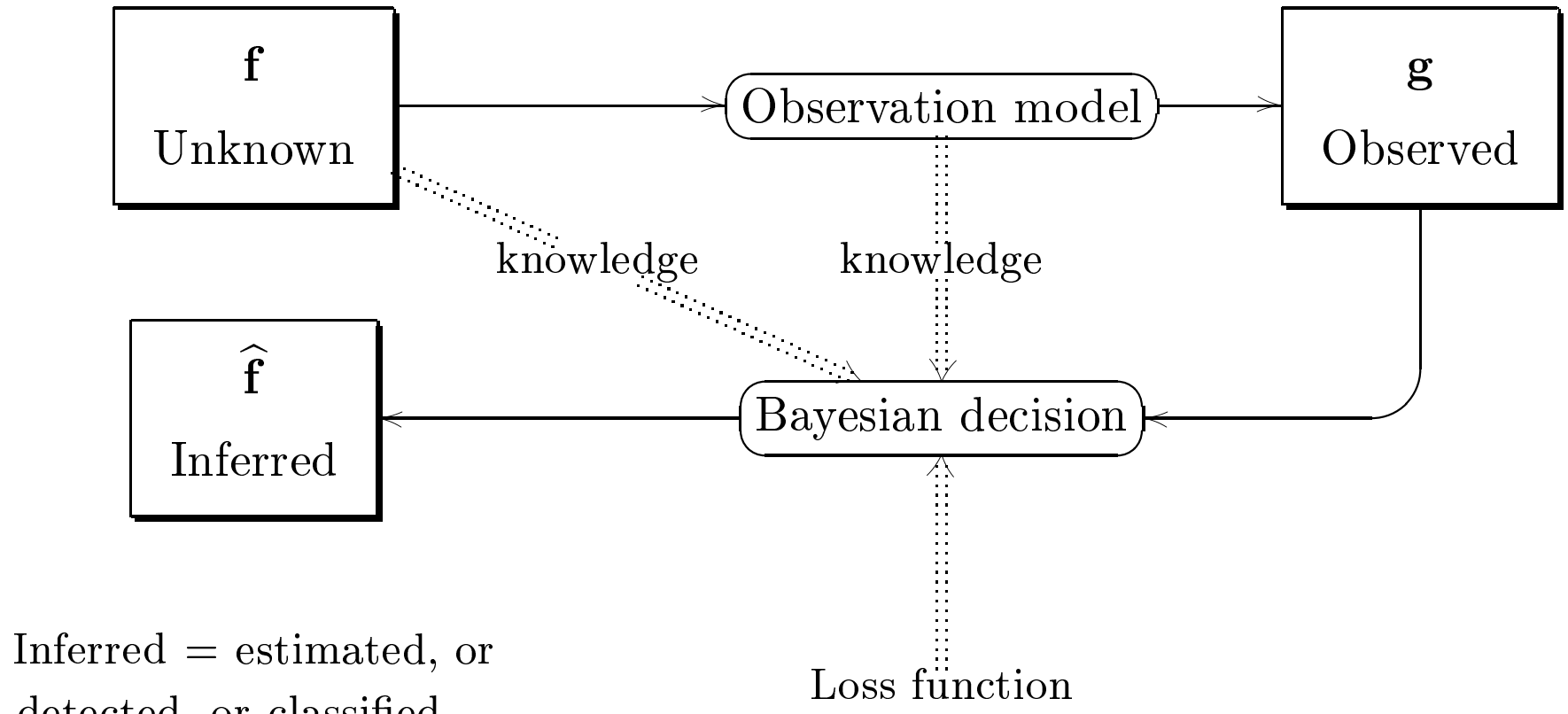
- They can not be solved without using *a priori* knowledge.
- Both $\mathbf{f}$ and $\mathbf{g}$ are high-dimensional.

  (e.g., images).

- They are naturally formulated as statistical inference problems.

Introduction to Bayesian theory

Basically, the Bayesian approach provides a way to "invert" an observation model, taking prior knowledge into account.



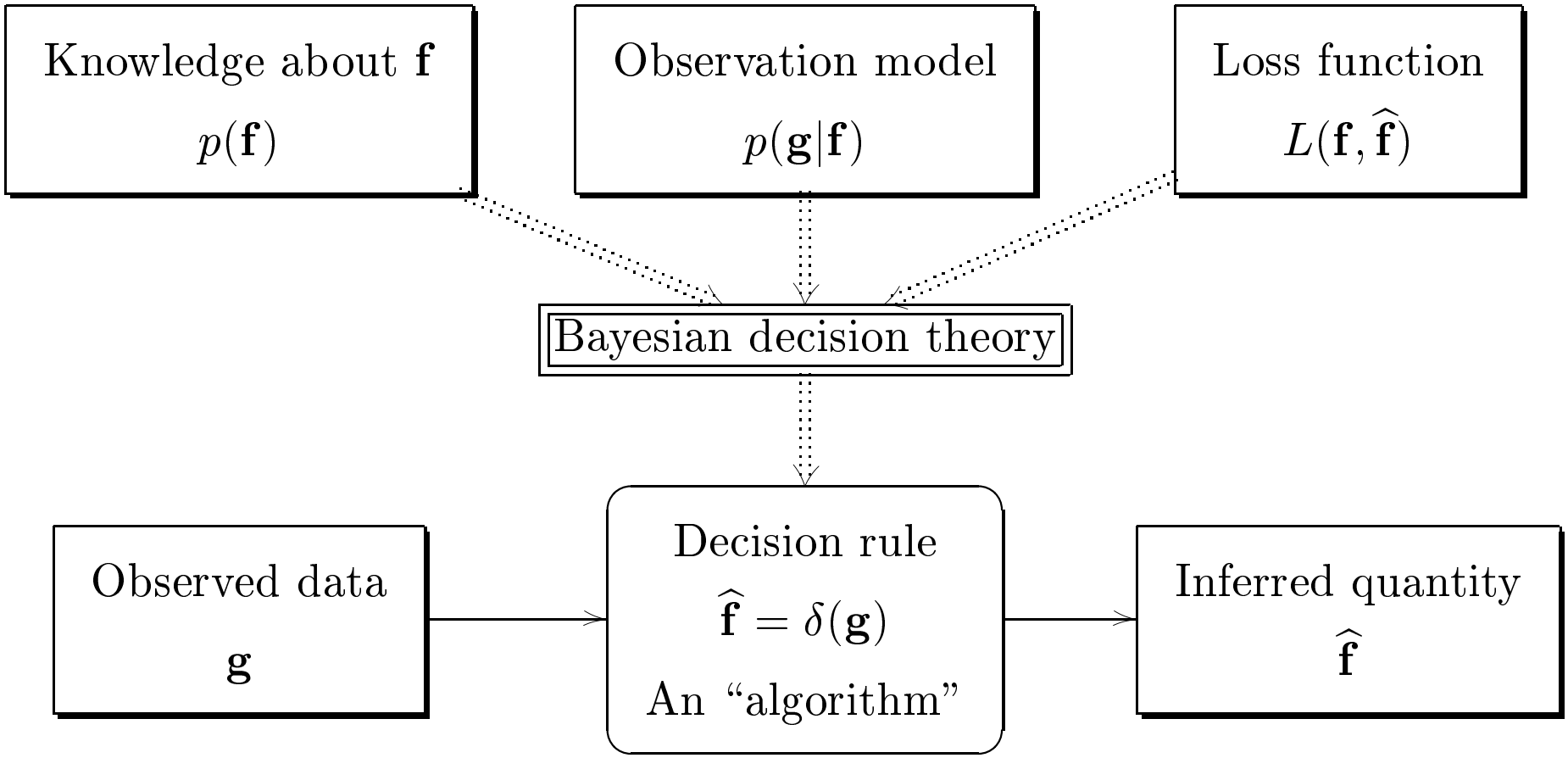
The Bayesian philosophy

Knowledge $\Leftrightarrow$ probability

- A subjective (non-frequentist) interpretation of probability.
- Probabilities express "degrees of belief".
- Example: *"there is a 20% probability that a certain patient has a tumor"*. Since we are considering one particular patient, this statement has no frequential meaning; it expresses a *degree of belief*.
- It can be shown that probability theory is **the** right tool to formally deal with "degrees of belief" or "knowledge";

  Cox (46), Savage (54), Good (60), Jeffreys (39, 61), Jaynes (63, 68, 91).

Bayesian decision theory



How are Bayesian decision rules derived?

By applying the fundamental principles of the Bayesian philosophy:

- Knowledge is expressed via probability functions.
- The "conditionality principle": any inference must be based (conditioned) on the observed data ($\mathbf{g}$).
- The "likelihood principle": The information contained in the observation $\mathbf{g}$ can only be carried via the likelihood function $p(\mathbf{f}|\mathbf{g})$.

Accordingly, knowledge about $\mathbf{f}$, once $\mathbf{g}$ is observed, is expressed by the *a posteriori* (or posterior) probability function:

$$\boxed{p(\mathbf{f}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})}{p(\mathbf{g})}}$$

"Bayes law"

How are Bayesian decision rules derived? (cont.)

- Once $\mathbf{g}$ is observed, knowledge about $\mathbf{f}$ is expressed by $p(\mathbf{f}|\mathbf{g})$.
- Given $\mathbf{g}$, what is the expected value of the loss function $L(\mathbf{f}, \widehat{\mathbf{f}})$?

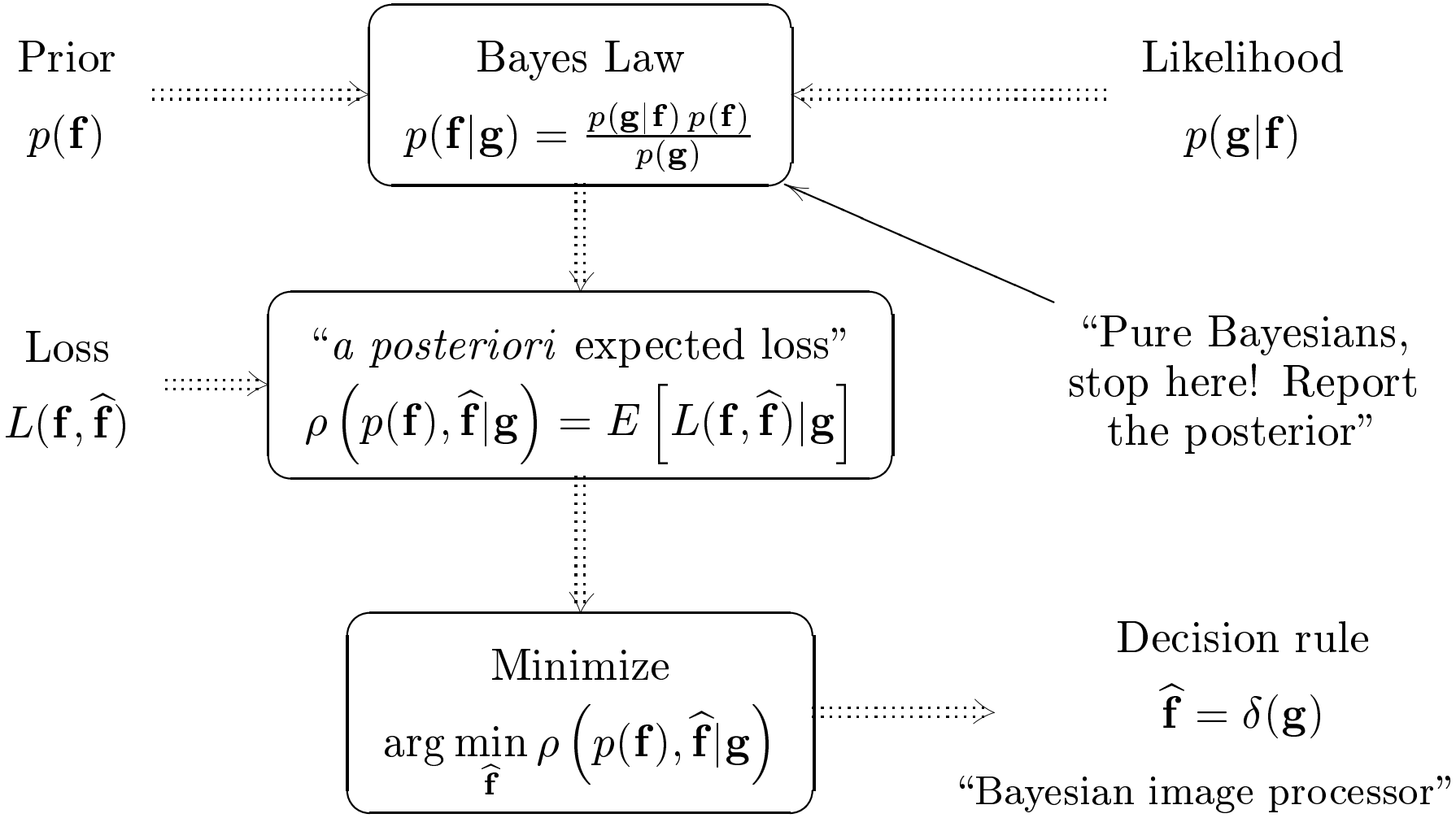
$$E\left[L(\mathbf{f}, \widehat{\mathbf{f}})|\mathbf{g}\right] = \int L(\mathbf{f}, \widehat{\mathbf{f}})\, p(\mathbf{f}|\mathbf{g})\, d\mathbf{f} \equiv \rho\left(p(\mathbf{f}), \widehat{\mathbf{f}}|\mathbf{g}\right)$$

...the so-called "*a posteriori* expected loss".

- An "optimal Bayes rule", is one minimizing $\rho\left(p(\mathbf{f}), \widehat{\mathbf{f}}|\mathbf{g}\right)$:

$$\widehat{\mathbf{f}}_{\text{Bayes}} = \delta_{\text{Bayes}}(\mathbf{g}) = \arg\min_{\widehat{\mathbf{f}}} \rho\left(\mathbf{p}(\mathbf{f}), \widehat{\mathbf{f}}|\mathbf{g}\right)$$

How are Bayesian decision rules derived? (cont.)



More on Bayes law.

$$p(\mathbf{f}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})}{p(\mathbf{g})}$$

- The numerator is the joint probability of $\mathbf{f}$ and $\mathbf{g}$:

$$p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f}) = p(\mathbf{g}, \mathbf{f}).$$

- The denominator is simply a normalizing constant,

$$p(\mathbf{g}) = \int p(\mathbf{g}, \mathbf{f})\, d\mathbf{f} = \int p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})\, d\mathbf{f}$$

  ...it is a marginal probability function.

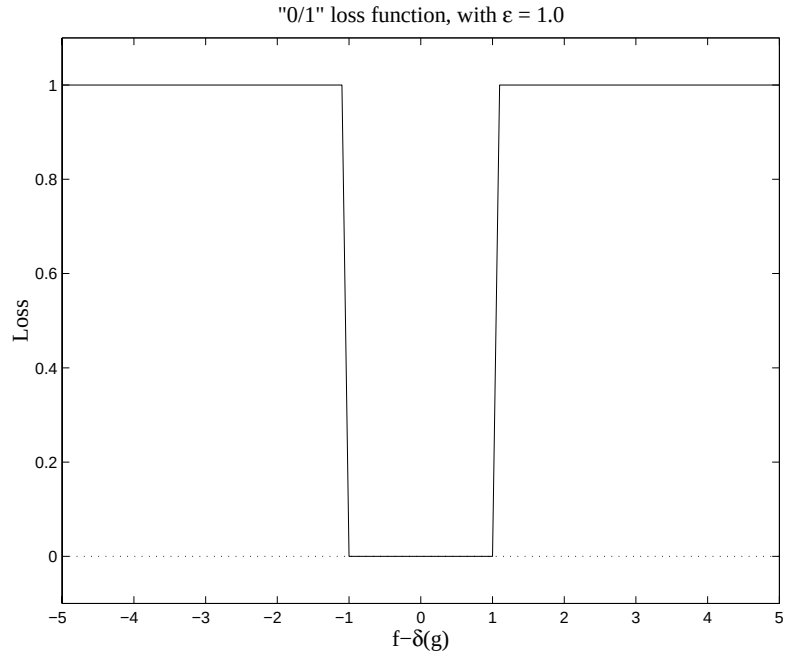
  Other names: unconditional, predictive, evidence.

- In discrete cases, rather than integral we have a summation.

The "0/1" loss function

For a scalar continuous $f \in \mathcal{F}$, e.g., $\mathcal{F} = I\!R$,

$$L_\varepsilon(f, \widehat{f}) = \begin{cases} 1 & \Leftarrow \quad |f - \widehat{f}| \geq \varepsilon \\ 0 & \Leftarrow \quad |f - \widehat{f}| < \varepsilon \end{cases}$$



The "0/1" loss function (cont.)

- Minimizing the "a posteriori" expected loss:

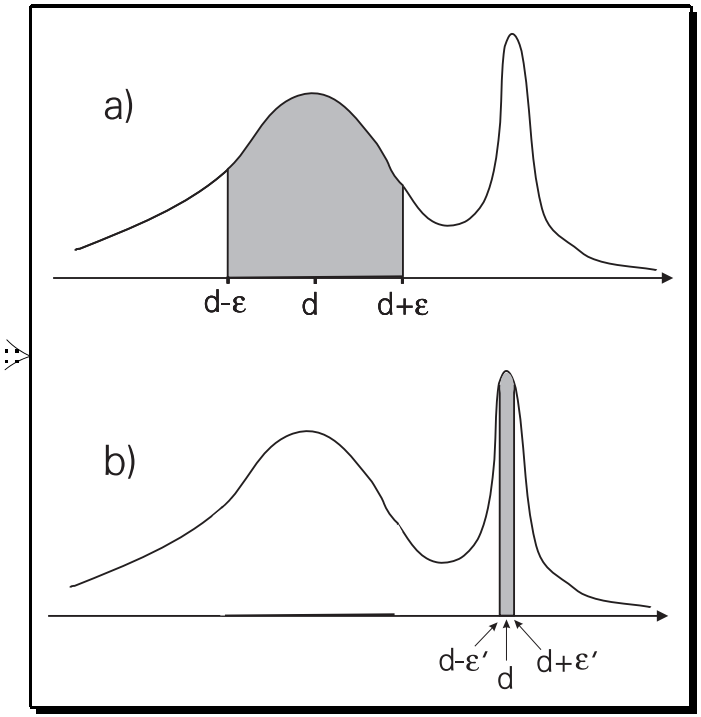
$$\begin{aligned}
\delta_\varepsilon(\mathbf{g}) &= \arg\min_d \int_{\mathcal{F}} L_\varepsilon(f,d)\, p(f|\mathbf{g})\, df \\
&= \arg\min_d \int_{f:|f-d|\geq\varepsilon} p(f|\mathbf{g})\, df \\
&= \arg\min_d \left(1 - \int_{f:|f-d|<\varepsilon} p(f|\mathbf{g})\, df\right) \\
&= \arg\max_d \int_{d-\varepsilon}^{d+\varepsilon} p(f|\mathbf{g})\, df
\end{aligned}$$

- Letting ε approach zero,

$$\begin{aligned} \lim_{\varepsilon \to 0} \delta_\varepsilon(\mathbf{g}) &= \lim_{\varepsilon \to 0} \arg\max_d \int_{d-\varepsilon}^{d+\varepsilon} p(f|\mathbf{g})\, df \\ &= \arg\max_f p(f|\mathbf{g}) \equiv \delta_{\text{MAP}}(\mathbf{g}) \equiv \widehat{f}_{\text{MAP}} \end{aligned}$$

...called the "maximum *a posteriori*" (MAP) estimator.

With ε decreasing, $\delta_\varepsilon(\mathbf{g})$ "looks for" the highest mode of $p(f|\mathbf{g})$



The "0/1" loss for a scalar discrete $f \in \mathcal{F}$

$$L(f, \widehat{f}) = \begin{cases} 1 & \Leftarrow \quad f \neq \widehat{f} \\ 0 & \Leftarrow \quad f = \widehat{f} \end{cases}$$

- Again, minimizing the "a posteriori" expected loss:

$$\begin{aligned} \delta(\mathbf{g}) &= \arg\min_d \sum_{f \in \mathcal{F}} L_\varepsilon(f - d)\, p(f|\mathbf{g}) \\ &= \arg\min_d \sum_{f \neq d} p(f|\mathbf{g}) \\ &= \arg\min_d \{-p(d|\mathbf{g}) + \underbrace{\sum_{f \in \mathcal{F}} p(f|\mathbf{g})}_{1}\} \\ &= \arg\max_f p(f|\mathbf{g}) \;\equiv\; \delta_{\text{MAP}}(\mathbf{g}) \;\equiv\; \widehat{f}_{\text{MAP}} \end{aligned}$$

...the "maximum *a posteriori*" (MAP) classifier/detector.

"Quadratic" loss function

For a scalar continuous $f \in \mathcal{F}$, e.g., $\mathcal{F} = I\!R$,

$$L\left(f, \widehat{f}\right) = \left(f - \widehat{f}\right)^2$$

- Minimizing the *a posteriori* expected loss,

$$\begin{aligned}
\delta_{\text{PM}}(\mathbf{g}) &= \arg\min_d E\left[(f-d)^2 \,|\mathbf{g}\right] \\
&= \arg\min_d \{\underbrace{E\left[f^2|\mathbf{g}\right]}_{Constant} + d^2 - 2\,d\,E\left[f|\mathbf{g}\right]\} \\
&= E\left[f|\mathbf{g}\right] \equiv \widehat{f}_{\text{PM}}
\end{aligned}$$

...the "posterior mean" (PM) estimator.

Example: Gaussian observations with a Gaussian prior.

- The observation model is

$$\begin{aligned} p(\mathbf{g}|f) &= p([g_1\ g_2\ ...g_n]^T|f) \sim \mathcal{N}([f\ f\ ...f]^T, \sigma^2\mathbf{I}) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(g_i - f)^2\right\} \end{aligned}$$

  where $\mathbf{I}$ denotes an identity matrix.

- The prior is

$$p(f) = \left(2\pi\phi^2\right)^{-1/2} \exp\left\{-\frac{f^2}{2\phi^2}\right\} \sim \mathcal{N}(0, \phi^2)$$

- From these two models, the posterior is simply

$$p(f|\mathbf{g}) \sim \mathcal{N}\left(\frac{\bar{\mathbf{g}}\phi^2}{\frac{\sigma^2}{n} + \phi^2}, \left(\frac{n}{\sigma^2} + \frac{1}{\phi^2}\right)^{-1}\right) \quad \text{with } \bar{\mathbf{g}} = \frac{g_1 + ... + g_n}{n}$$

Example: Gaussian observations with a Gaussian prior (cont.).

- As seen in the previous slide

$$p(f|\mathbf{g}) \sim \mathcal{N}\left(\bar{\mathbf{g}}\frac{\phi^2}{\frac{\sigma^2}{n}+\phi^2}, \left(\frac{n}{\sigma^2}+\frac{1}{\phi^2}\right)^{-1}\right)$$

- Then, since the mean and the mode of a Gaussian coincide,

$$\widehat{f}_{\text{MAP}} = \widehat{f}_{\text{PM}} = \bar{\mathbf{g}}\frac{\phi^2}{\frac{\sigma^2}{n}+\phi^2};$$

  the estimate is a "shrunken" version of the sample mean $\bar{\mathbf{g}}$.

- If the prior had mean μ, we would have

$$\widehat{f}_{\text{MAP}} = \widehat{f}_{\text{PM}} = \frac{\mu\frac{\sigma^2}{n}+\bar{\mathbf{g}}\phi^2}{\frac{\sigma^2}{n}+\phi^2};$$

  i.e., the estimate is a weighted average of μ and $\bar{\mathbf{g}}$

Example: Gaussian observations with a Gaussian prior (cont.).

- Observe that

$$\lim_{n\to\infty} \frac{\mu\frac{\sigma^2}{n} + \bar{\mathbf{g}}\phi^2}{\frac{\sigma^2}{n} + \phi^2} = \lim_{n\to\infty} \frac{\mu\sigma^2 + n\bar{\mathbf{g}}\phi^2}{\sigma^2 + n\phi^2} = \bar{\mathbf{g}}$$

  i.e., as n increases, the data dominates the estimate.

- The posterior variance does not depend on $\mathbf{g}$,

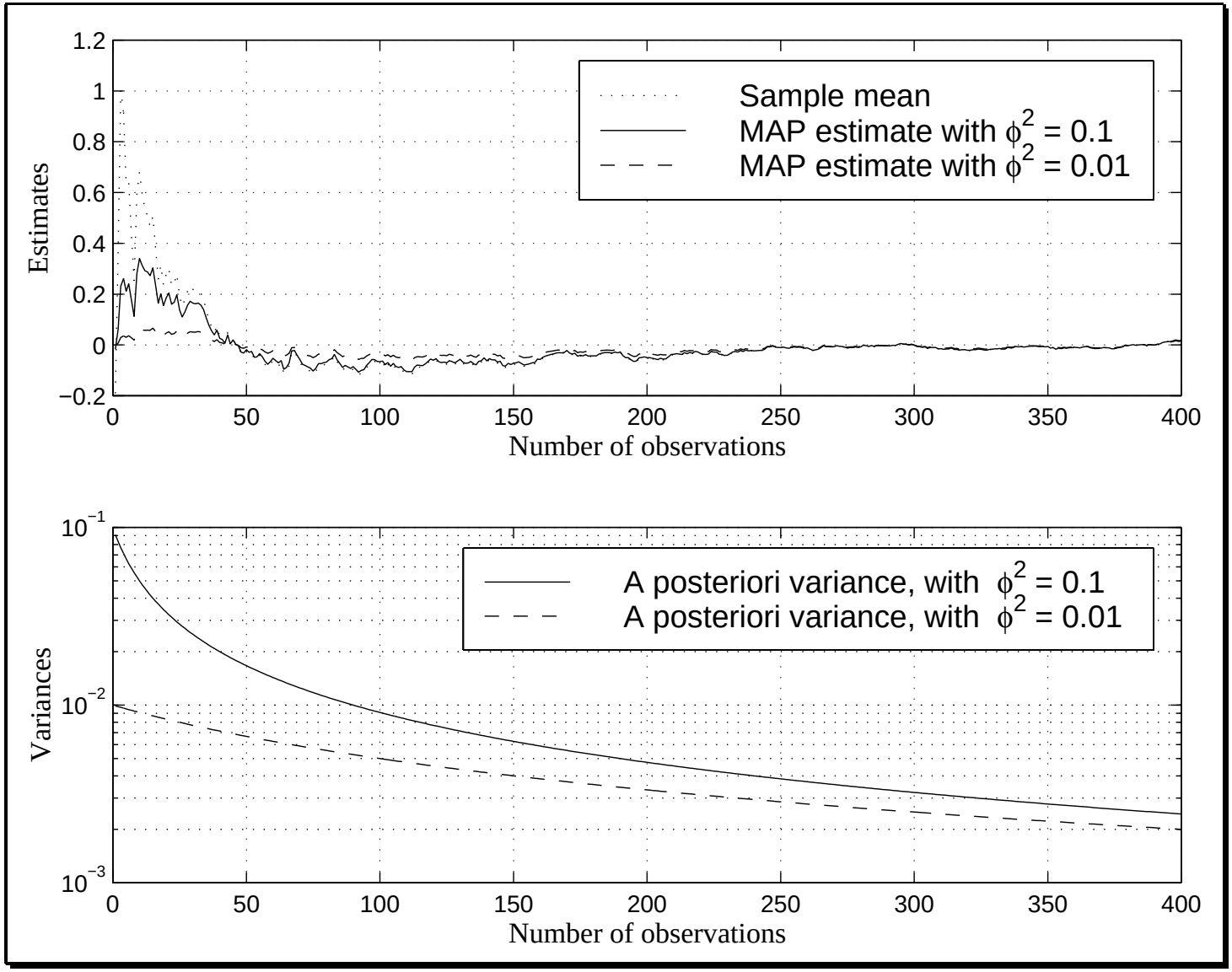
$$E\left[(f - \widehat{f})^2 | \mathbf{g}\right] = \left(\frac{n}{\sigma^2} + \frac{1}{\phi^2}\right)^{-1},$$

  inversely proportional to the degree of confidence on the estimate.

- Notice also that

$$\lim_{n\to\infty} E\left[(f - \widehat{f})^2 | \mathbf{g}\right] = \lim_{n\to\infty} \left(\frac{n}{\sigma^2} + \frac{1}{\phi^2}\right)^{-1} = 0;$$

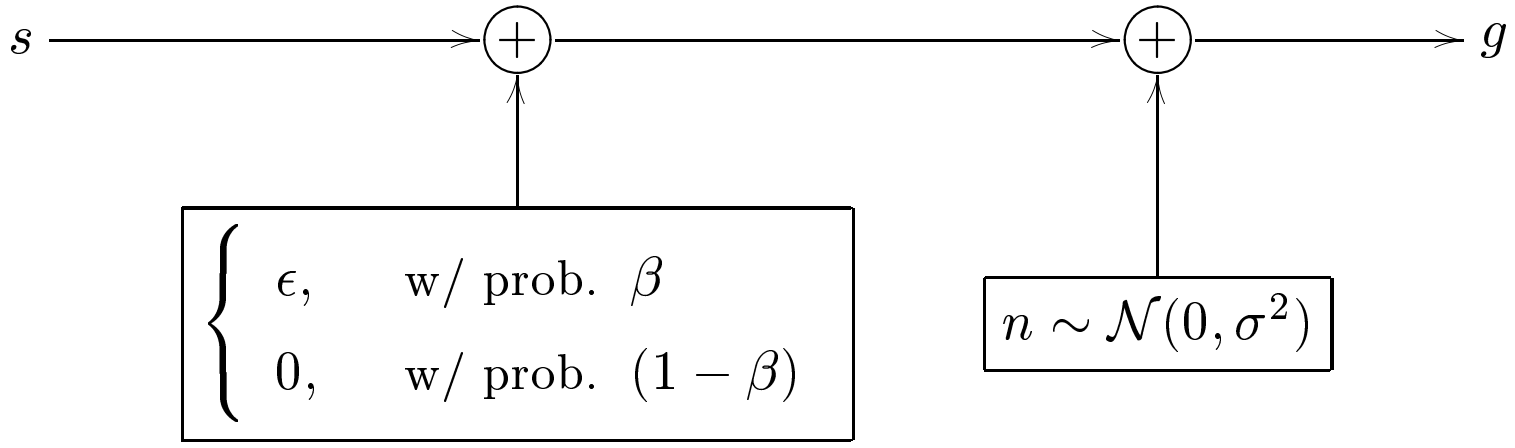
  ...as $n \to \infty$ the confidence on the estimate becomes absolute.


Sample mean
MAP estimate with $\phi^2 = 0.1$
MAP estimate with $\phi^2 = 0.01$
Estimates
Number of observations
A posteriori variance, with $\phi^2 = 0.1$
A posteriori variance, with $\phi^2 = 0.01$
Variances
Number of observations

Example: Gaussian mixture observations with a Gaussian prior.

- "Mixture" observation model

$$p(g|s) = \frac{\beta}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(g-s-\epsilon)^2}{2\sigma^2}\right\} + \frac{1-\beta}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(g-s)^2}{2\sigma^2}\right\},$$



- Gaussian prior $p(s) \sim \mathcal{N}(0, \phi^2)$.

Example: Gaussian mixture observations with a Gaussian prior (cont.).

The posterior:

$$p(s|g) \propto \beta \exp\left\{-\frac{(g-s-\epsilon)^2}{2\sigma^2} - \frac{s^2}{2\phi^2}\right\} + (1-\beta)\exp\left\{-\frac{(g-s)^2}{2\sigma^2} - \frac{s^2}{2\phi^2}\right\}$$

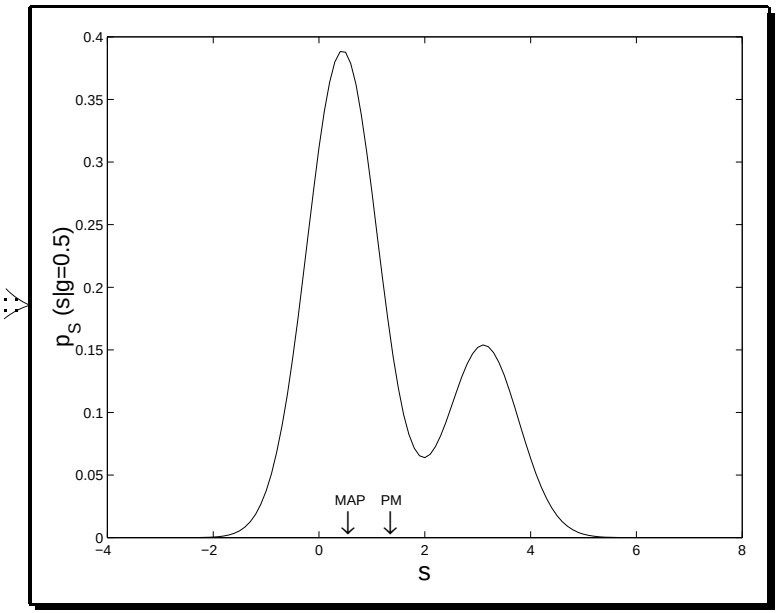
Example:

$\beta = 0.6$

$\phi^2 = 4$

$\sigma^2 = 0.5$

$g = 0.5$



PM = "compromise"; MAP = largest mode.

Improper priors and "maximum likelihood" inference

- Recall that the posterior is computed according to

$$p(\mathbf{f}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})}{p(\mathbf{g})}$$

- If the MAP criterion is being used, and $p(\mathbf{f}) = k$,

$$\begin{aligned} \widehat{f}_{\text{MAP}} &= \arg\max_{\mathbf{f}} \frac{p(\mathbf{g}|\mathbf{f})\, k}{k \int p(\mathbf{g}|\mathbf{f})\, d\mathbf{f}} \\ &= \arg\max_{\mathbf{f}} p(\mathbf{g}|\mathbf{f}), \end{aligned}$$

  ...the "maximum likelihood" (ML) estimate.

- In the discrete case, simply replace the integral by a summation.

Improper priors and maximum likelihood inference (cont.)

- If the space to which $\mathbf{f}$ belongs is unbounded, e.g., $\mathbf{f} \in I\!R^m$, or $f \in I\!N$, the prior is "improper":

$$\int p(\mathbf{f})\, d\mathbf{f} = \int k\, d\mathbf{f} = \infty.$$

  or

$$\sum p(\mathbf{f}) = k \sum 1 = \infty.$$

- If the posterior is proper, all the estimates are still well defined.
- Improper priors reinforce the "knowledge" interpretation of probabilities.

Compound inference: Inferring a set of unknowns

- Now, $\mathbf{f}$ is a (say, m-dimensional) vector,

$$\mathbf{f} = [f_1, f_2, ..., f_m]^T.$$

- Loss functions for compound problems:

  Additive: Such that $L(\mathbf{f}, \widehat{\mathbf{f}}) = \sum_{i=1}^{M} L_i(f_i, \widehat{f}_i)$.

  Non-additive: This decomposition does not exist.

- Optimal Bayes rules are still

$$\widehat{\mathbf{f}}_{\text{Bayes}} = \delta_{\text{Bayes}}(\mathbf{g}) = \arg\min_{\widehat{\mathbf{f}}} \int \mathbf{L}(\mathbf{f}, \widehat{\mathbf{f}}) \mathbf{p}(\mathbf{f}|\mathbf{g}) \, \mathbf{df}$$

Compound inference with non-additive loss functions.

There is nothing fundamentally new in this case.

- The "0/1" loss, for a vector $\mathbf{f}$ (e.g., $\mathcal{F} = I\!R^m$):

$$L_\varepsilon(\mathbf{f}, \widehat{\mathbf{f}}) = \begin{cases} 1 & \Leftarrow \quad \| \mathbf{f} - \widehat{\mathbf{f}} \| \geq \varepsilon \\ 0 & \Leftarrow \quad \| \mathbf{f} - \widehat{\mathbf{f}} \| < \varepsilon \end{cases}$$

- Following the same derivation yields

$$\widehat{\mathbf{f}}_{\text{MAP}} = \delta_{\text{MAP}}(\mathbf{g}) = \arg \max_{\mathbf{f}} p(\mathbf{f}|\mathbf{g})$$

  i.e., the MAP estimate is the joint mode of the *a posteriori* probability function.

- Exactly the same expression is obtained for discrete problems.

Compound inference with non-additive loss functions (cont.)

- The **quadratic** loss, for $\mathbf{f} \in I\!R^m$: $L(\mathbf{f}, \widehat{\mathbf{f}}) = (\mathbf{f} - \widehat{\mathbf{f}})^T \mathbf{Q} (\mathbf{f} - \widehat{\mathbf{f}})$ where $\mathbf{Q}$ is a symmetric positive-definite $(m \times m)$ matrix.

- Minimizing the *a posteriori* expected loss,

$$\begin{aligned}
\delta_{\text{PM}}(\mathbf{g}) &= \arg\min_{\widehat{\mathbf{f}}} E\left[(\mathbf{f} - \widehat{\mathbf{f}})^T \mathbf{Q} (\mathbf{f} - \widehat{\mathbf{f}}) | \mathbf{g}\right] \\
&= \arg\min_{\widehat{\mathbf{f}}} \{ \underbrace{E\left[\mathbf{f}^T \mathbf{Q} \mathbf{f} | \mathbf{g}\right]}_{Constant} + \widehat{\mathbf{f}}^T \mathbf{Q} \widehat{\mathbf{f}} - 2\, \widehat{\mathbf{f}}^T \mathbf{Q}\, E\left[\mathbf{f}|\mathbf{g}\right] \} \\
&= \text{solution of } \left\{ \mathbf{Q}\widehat{\mathbf{f}} = \mathbf{Q} E\left[\mathbf{f}|\mathbf{g}\right] \right\} \qquad (\mathbf{Q} \text{ has inverse}) \\
&= E\left[\mathbf{f}|\mathbf{g}\right] \equiv \widehat{f}_{\text{PM}}
\end{aligned}$$

  ...still the "posterior mean" (PM) estimator.

- Remarkably, this is true for any symmetric positive-definite $\mathbf{Q}$. Special case: $\mathbf{Q}$ is diagonal, the loss function is additive.

Compound inference with additive loss functions

- Recall that, in this case, $L(\mathbf{f}, \widehat{\mathbf{f}}) = \sum_{i=1}^{M} L_i(f_i, \widehat{f}_i)$.
- The optimal Bayes rule

$$\begin{aligned}
\delta(\mathbf{g})_{\text{Bayes}} &= \arg\min_{\widehat{\mathbf{f}}} \int \underbrace{\sum_{i=1}^{m} L_i(f_i, \widehat{f}_i)}_{L(\mathbf{f},\widehat{\mathbf{f}})} p(\mathbf{f}|\mathbf{g})\, d\mathbf{f} \\
&= \arg\min_{\widehat{\mathbf{f}}} \sum_{i=1}^{m} \int L_i(f_i, \widehat{f}_i)\, p(\mathbf{f}|\mathbf{g})\, d\mathbf{f} \\
&= \arg\min_{\widehat{\mathbf{f}}} \sum_{i=1}^{m} \int L_i(f_i, \widehat{f}_i) \left(\int p(\mathbf{f}|\mathbf{g})\, d\mathbf{f}_{-i} \right) df_i
\end{aligned}$$

where $d\mathbf{f}_{-i}$ denotes $df_1 ... df_{i-1} df_{i+1} ... df_m$, that is, integration with respect to all variables except f_i

Compound inference with additive loss functions (cont.)

- From the previous slide:

$$\delta(\mathbf{g})_{\text{Bayes}} = \arg\min_{\widehat{\mathbf{f}}} \sum_{i=1}^{m} \int L_i(f_i, \widehat{f}_i) \left(\int p(\mathbf{f}|\mathbf{g}) \, d\mathbf{f}_{-i} \right) df_i$$

- But, $\int p(\mathbf{f}|\mathbf{g}) \, d\mathbf{f}_{-i} = p(f_i|\mathbf{g})$,
  the *a posteriori* marginal of variable f_i.

- Then, $\delta(\mathbf{g})_{\text{Bayes}} = \arg\min_{\widehat{\mathbf{f}}} \sum_{i=1}^{m} \int L_i(f_i, \widehat{f}_i) p(f_i|\mathbf{g}) \, df_i$,
  that is, $\widehat{f}_{i\,\text{Bayes}} = \arg\min_{\widehat{f}_i} \int L_i(f_i, \widehat{f}_i) p(f_i|\mathbf{g}) \, df_i \quad i = 1, 2, ...m$

- Conclusion: each estimate is the minimizer of the corresponding marginal *a posteriori* expected loss

Additive loss functions: Special cases

- The additive "0/1" loss function: $L(\mathbf{f}, \widehat{\mathbf{f}}) = \sum_{i=1}^{m} L_i(f_i, \widehat{f}_i)$,

  where each $L_i(f_i, \widehat{f}_i)$ is a "0/1" loss function for scalar arguments.

  According to the general result,

  $$\widehat{\mathbf{f}}_{\text{MPM}} = \left[\arg\max_{f_1} p(f_1|\mathbf{g}) \quad \arg\max_{f_2} p(f_2|\mathbf{g}) \quad \cdots \arg\max_{f_m} p(f_m|\mathbf{g}) \right]^T$$

  the *maximizer of posterior marginals* (MPM).

- The additive quadratic loss function:
  $L(\mathbf{f}, \widehat{\mathbf{f}}) = \sum_{i=1}^{m} (f_i - \widehat{f}_i)^2 = (\mathbf{f} - \widehat{\mathbf{f}})^T (\mathbf{f} - \widehat{\mathbf{f}})$.

  The general result for quadratic loss functions is still valid.

  This is a natural fact because the mean is intrinsically marginal.

Example: Gaussian observations and Gaussian prior.

- Observation model: linear operator (matrix) plus additive white Gaussian noise:

$$\mathbf{g} = \mathbf{H}\,\mathbf{f} + \mathbf{n}, \qquad \text{where} \quad \mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$$

- Corresponding likelihood function

$$p(\mathbf{g}|\mathbf{f}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \parallel \mathbf{H}\,\mathbf{f} - \mathbf{g} \parallel^2 \right\}$$

- Gaussian prior:

$$p(\mathbf{f}) = \frac{(2\pi)^{-n/2}}{\sqrt{\det(\mathbf{K})}} \exp\left\{ -\frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} \right\}$$

Example: Gaussian observations and Gaussian prior (cont.)

- The *a posteriori* (joint) probability density function is still Gaussian

$$p(\mathbf{f}|\mathbf{g}) \sim \mathcal{N}\left(\widehat{\mathbf{f}}, \mathbf{P}\right);$$

with $\widehat{\mathbf{f}}$ being the MAP and PM estimate, given by

$$\begin{aligned} \widehat{\mathbf{f}} &= \arg\min_{\mathbf{f}} \left\{ \mathbf{f}^T \left[\sigma^2 \mathbf{K}^{-1} + \mathbf{H}^T\mathbf{H}\right] \mathbf{f} - 2\mathbf{f}^T\mathbf{H}^T\mathbf{g} \right\} \\ &= \left[\sigma^2 \mathbf{K}^{-1} + \mathbf{H}^T\mathbf{H}\right]^{-1} \mathbf{H}^T\mathbf{g}. \end{aligned}$$

- This is also called the (vector) Wiener filter.

Example: Gaussian observations and Gaussian prior; special cases.

No noise: Absence of noise $\Leftrightarrow$ $\sigma^2 = 0$

$$\begin{aligned} \widehat{\mathbf{f}} &= \left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{g}. \\ &= \arg\min_{\mathbf{f}} \left\{ \|\mathbf{H}\mathbf{f} - \mathbf{g}\|^2 \right\} \end{aligned}$$

- $\left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T \equiv \mathbf{H}^\dagger$ is called the Moore-Penrose *pseudo* (or *generalized*) *inverse* of matrix $\mathbf{H}$.
- If $\mathbf{H}^{-1}$ exists, $\mathbf{H}^\dagger = \mathbf{H}^{-1}$;
- If $\mathbf{H}$ is not invertible, $\mathbf{H}^\dagger$ provides its *least-squares* sense pseudo-solution.
- This estimate is also the *maximum likelihood* one.

Example: Gaussian observations and Gaussian prior; special cases.

Prior covariance up to a factor: $\mathbf{K} = \phi^2 \mathbf{B}$; diagonal elements of $\mathbf{B}$ equal to 1. ϕ^2 can be seen as the "prior variance".

- $\mathbf{K}^{-1} = \mathbf{B}^{-1}/\phi^2$ is positive definite $\Rightarrow$ exists unique symmetric $\mathbf{D}$ such that $\mathbf{DD} = \mathbf{D}^T\mathbf{D} = \mathbf{B}^{-1}$.
- This allows writing $$\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} \left\{ \|\mathbf{g} - \mathbf{H}\mathbf{f}\|^2 + \frac{\sigma^2}{\phi^2} \|\mathbf{D}\mathbf{f}\|^2 \right\}$$
- In regularization theory parlance, $\|\mathbf{D}\mathbf{f}\|^2$ is called the regularizing term, and σ^2/ϕ^2 the regularization parameter.
- We can also write
$$\widehat{\mathbf{f}} = \left(\frac{\sigma^2}{\phi^2} \mathbf{B}^{-1} + \mathbf{H}^T\mathbf{H} \right)^{-1} \mathbf{H}^T\mathbf{g};$$
σ^2/ϕ^2 controls the relative weight of the prior and the data.

Summary of what we have seen up to this point

- Image analysis problems are inference problems
- Introduction to Bayesian inference:
  - Fundamental principles: knowledge as probability, likelihood and conditionality.
  - Fundamental tool: Bayes rule.
  - Necessary models: observation model, prior, loss function.
  - *A posteriori* expected loss and optimal Bayes rules.
  - The "0/1" loss function and MAP inference.
  - The quadratic error loss function and posterior mean estimation.
  - Example: Gaussian observations and Gaussian prior.
  - Example: Mixture of Gaussians observations and Gaussian prior.
  - Improper priors and *maximum likelihood* (ML) inference.
  - Compound inference: additive and non-additive loss functions.
  - Example: Gaussian observations with Gaussian prior.

Conjugate priors: Looking for computational convenience

- Sometimes the prior knowledge is vague enough to allow tractability concerns to come into play.
- In other words: choose priors compatible with knowledge, but leading to a tractable *a posteriori* probability function.
- *Conjugate priors* formalize this goal.
- A family of likelihood functions $\mathcal{L} = \{p(\mathbf{g}|\mathbf{f}),\ \ \mathbf{f} \in \mathcal{F}\}$
- A (parametrized) family of priors $\mathcal{P} = \{p(\mathbf{f}|\boldsymbol{\theta}),\ \ \boldsymbol{\theta} \in \Theta\}$
- $\mathcal{P}$ is a *conjugate family* for $\mathcal{L}$, if

$$\left\{ \begin{array}{lcl} p(\mathbf{g}|\mathbf{f}) & \in & \mathcal{L} \\ p(\mathbf{f}|\boldsymbol{\theta}) & \in & \mathcal{P} \end{array} \right\} \Rightarrow p(\mathbf{f}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f}|\boldsymbol{\theta})}{p(\mathbf{g})} \in \mathcal{P}$$

  i.e., $\exists\, \boldsymbol{\theta}' \in \Theta$, such that $p(\mathbf{f}|\mathbf{g}) = p(\mathbf{f}|\boldsymbol{\theta}')$.

Conjugate priors: A simple example

- The family of Gaussian likelihood functions of common variance

$$\mathcal{L} = \left\{ p(g|f) \sim \mathcal{N}(f, \sigma^2), \;\; f \in I\!R \right\}$$

- The family of Gaussian priors of arbitrary mean and variance

$$\mathcal{P} = \left\{ p(f|\mu, \phi^2) \sim \mathcal{N}(\mu, \phi^2), \;\; (\mu, \phi^2) \in I\!R \times I\!R^+ \right\}$$

- The *a posteriori* probability density function is

$$p(f|g) \sim \mathcal{N}\left(\frac{\mu\sigma^2 + g\phi^2}{\sigma^2 + \phi^2}, \frac{\sigma^2\phi^2}{\sigma^2 + \phi^2} \right) \in \mathcal{P}$$

- Very important: computing the *a posteriori* probability function only involves "updating" parameters of the prior.

Conjugate priors: Another example

- θ is the (unknown) "heads" probability of a given coin.
- Outcomes of a sequence of n tosses: $\mathbf{x} = (x_1, \ldots, x_n)$, $x_i \in \{1, 0\}$.
- Likelihood function (Bernoulli), with $n_h(\mathbf{x}) = x_1 + x_2 + ... + x_n$,

$$p(\mathbf{x}|\theta) = \theta^{n_h(\mathbf{x})} (1-\theta)^{n - n_h(\mathbf{x})}.$$

- *A priori* belief: "θ should be close to $1/2$".
- Conjugate prior: the Beta density

$$p(\theta|\alpha, \beta) \sim \mathcal{B}\mathrm{e}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1},$$

defined for $\theta \in [0, 1]$ and $\alpha, \beta > 0$.

Conjugate priors: Bernoulli example (cont.)

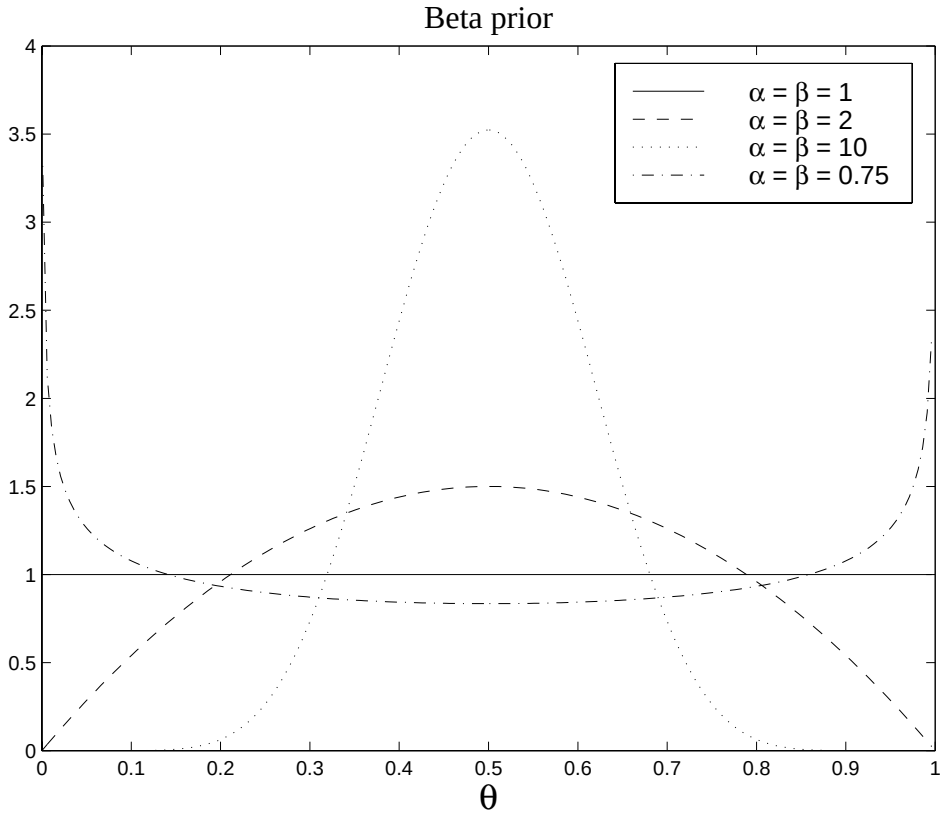
- Main features of $\mathcal{B}\text{e}(\alpha, \beta)$:

$$
\begin{aligned}
E[\theta|\alpha,\beta] &= \frac{\alpha}{\alpha+\beta} \qquad \text{(mean)} \\
E\left[\left(\theta - \frac{\alpha}{\alpha+\beta}\right)^2 \middle| \alpha,\beta\right] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \qquad \text{(variance)} \\
\arg\max_{\theta} p(\theta|\alpha,\beta) &= \frac{\alpha-1}{\alpha+\beta-2} \qquad \text{(mode, if } \alpha > 1\text{)},
\end{aligned}
$$

- "Pull" the estimate towards 1/2: choose $\alpha = \beta$.
- The quantity $\alpha = \beta$ controls "how strongly we pull".

Conjugate priors: Bernoulli example (cont.)

Several Beta densities:



For $\alpha = \beta \leq 1$, qualitatively different behavior: the mode at $1/2$ disappears.

Conjugate priors: Bernoulli example (cont.)

- The *a posteriori* distribution is again Beta

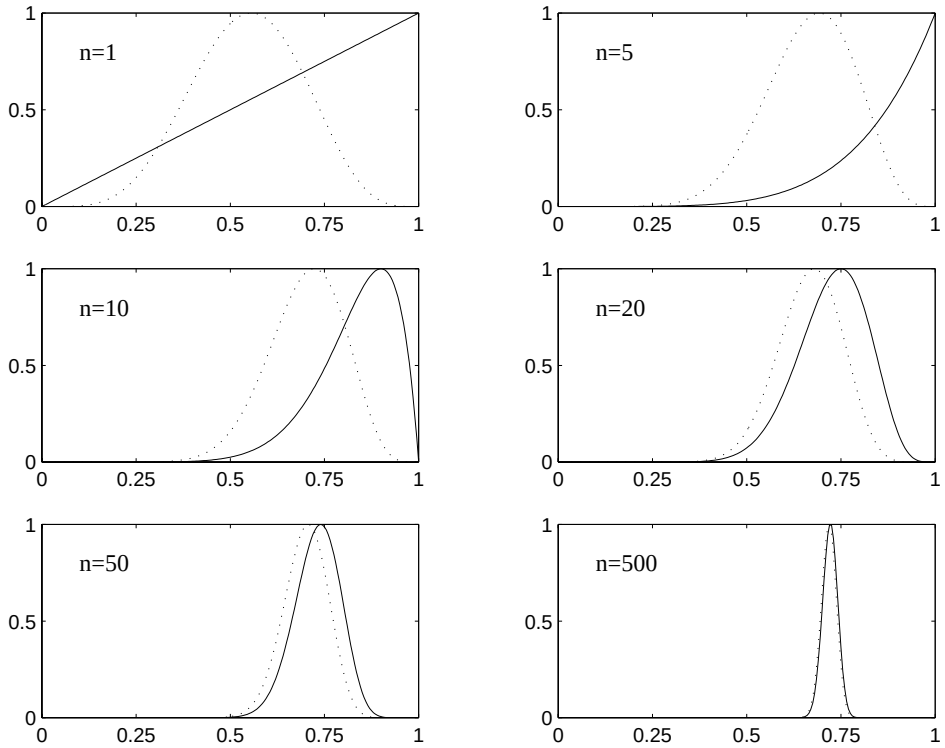
$$p(\theta|\mathbf{x}, \alpha, \beta) \sim \mathcal{B}\mathrm{e}\left(\alpha + n_h(\mathbf{x}), \beta + n - n_h(\mathbf{x})\right)$$

- Bayesian estimates of θ

$$\begin{aligned} \widehat{\theta}_{\mathrm{PM}} &= \delta_{\mathrm{PM}}(\mathbf{x}) = \frac{\alpha + n_h(\mathbf{x})}{\alpha + \beta + n} \\ \widehat{\theta}_{\mathrm{MAP}} &= \delta_{\mathrm{MAP}}(\mathbf{x}) = \frac{\alpha + n_h(\mathbf{x}) - 1}{\alpha + \beta + n - 2}. \end{aligned}$$

Conjugate priors: Bernoulli example (cont.)

Evolution of the *a posteriori* densities, for a $\mathcal{B}\mathrm{e}(5,5)$ prior (dotted line) and $\mathcal{B}\mathrm{e}(1,1)$ flat prior (solid line).



Conjugate priors: Variance of Gaussian observations

- n i.i.d. zero-mean Gaussian observations of unknown variance $\sigma^2 = 1/\theta$
- Likelihood function

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \sqrt{\frac{\theta}{2\pi}} \exp\left\{-\theta \frac{x_i^2}{2}\right\} = \left(\frac{\theta}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\theta}{2}\sum_{i=1}^{n} x_i^2\right\}.$$

- Conjugate prior: the Gamma density.

$$p(\theta|\alpha,\beta) \sim \mathcal{G}\mathrm{a}(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}\exp\{-\beta\theta\}$$

for $\theta \in [0,\infty)$ (recall $\theta = 1/\sigma^2$) and $\alpha, \beta > 0$.

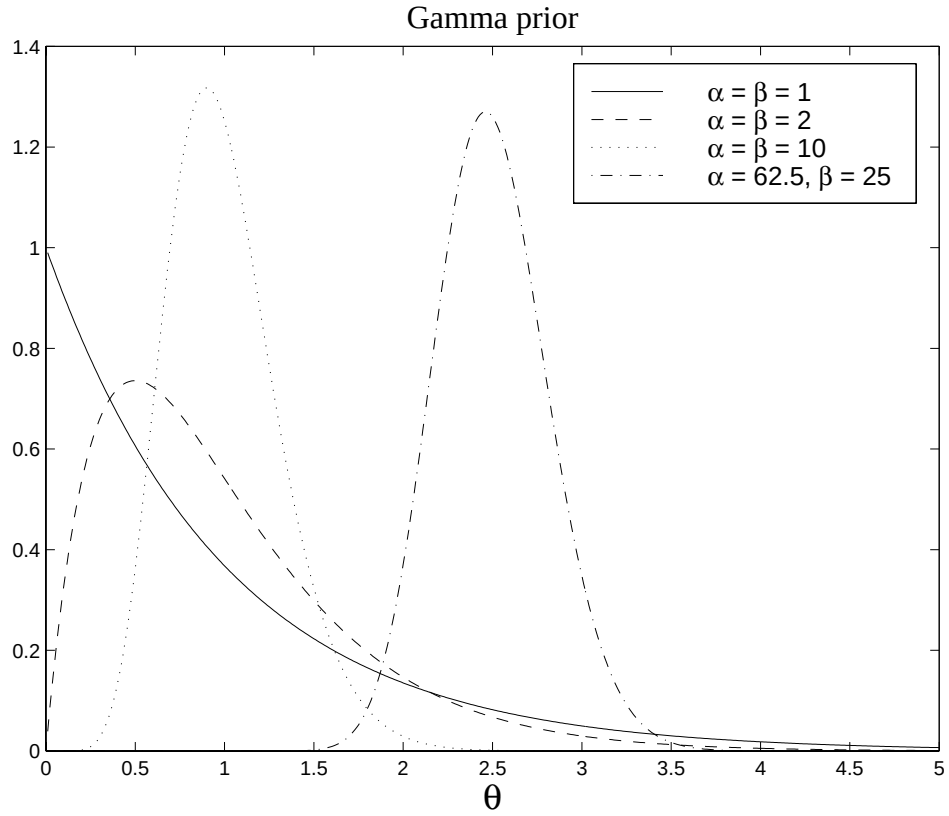
Conjugate priors: Variance of Gaussian observations (cont.)

- Main features of the Gamma density:

$$
\begin{aligned}
E[\theta|\alpha,\beta] &= \frac{\alpha}{\beta} \quad \text{(mean)} \\
E\left[\left(\theta-\frac{\alpha}{\beta}\right)^2\middle|\,\alpha,\beta\right] &= \frac{\alpha}{\beta^2} \quad \text{(variance)} \\
\arg\max_{\theta} p(\theta|\alpha,\beta) &= \frac{\alpha-1}{\beta} \quad \text{(mode, if } \alpha \geq 1\text{)},
\end{aligned}
$$

Conjugate priors: Variance of Gaussian observations (cont.)

Several Gamma densities:



Conjugate priors: Variance of Gaussian observations (cont.)

- *A posteriori* density:

$$p(\theta|x_1, x_2, ..., x_n) \sim \mathcal{G}\text{a}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n} x_i^2\right).$$

- The corresponding Bayesian estimates

$$\widehat{\theta}_{\text{PM}} = \left(\frac{2\alpha}{n} + 1\right)\left(\frac{2\beta}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{-1}$$

$$\widehat{\theta}_{\text{MAP}} = \left(\frac{2\alpha}{n} + 1 - \frac{2}{n}\right)\left(\frac{2\beta}{n} + \frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{-1}.$$

- Both estimates converge to the ML estimate:

$$\lim_{n\to\infty} \widehat{\theta}_{\text{PM}} = \lim_{n\to\infty} \widehat{\theta}_{\text{MAP}} = \widehat{\theta}_{\text{ML}} = n\left(\sum_{i=1}^{n} x_i^2\right)^{-1}$$

The von Mises Theorem

As long as the prior is continuous and not zero at the location of the ML estimate, then, the MAP estimate converges to the ML estimate as the number of data points n increases.

Bayesian model selection

- Scenario: there are K models available, i.e., $m \in \{m_1, ..., m_K\}$
- Given model m,

  Likelihood function: $p(\mathbf{g}|\mathbf{f}_{(m)}, m)$

  Prior: $p(\mathbf{f}_{(m)}|m)$

  Under different m's, $\mathbf{f}_{(m)}$ may have different meanings, and sizes.
- *A priori* model probabilities $\{p(m); m = m_1, ..., m_K\}$.
- The *a posteriori* probability function is

$$p(m, \mathbf{f}_{(m)}|\mathbf{g}) = \frac{p(\mathbf{g}|\mathbf{f}_{(m)}, m)\, p(\mathbf{f}_{(m)}, m)}{p(\mathbf{g})} = \frac{p(\mathbf{g}|\mathbf{f}_{(m)}, m)\, p(\mathbf{f}_{(m)}|m)\, p(m)}{p(\mathbf{g})}$$

- Seen strictly as a model selection problem, the natural loss function is the "0/1" with respect to the model, i.e.

$$L[(m, \mathbf{f}_{(m)}), (\widehat{m}, \widehat{\mathbf{f}}_{(\widehat{m})})] = \begin{cases} 0 & \Leftarrow \quad \widehat{m} = m \\ 1 & \Leftarrow \quad \widehat{m} \neq m \end{cases}$$

- The resulting rule is the "most probable mode *a posteriori*"

$$\begin{aligned} \widehat{m} &= \arg\max_m p(m|\mathbf{g}) = \arg\max_m \int p(m, \mathbf{f}_{(m)}|\mathbf{g})\, d\mathbf{f}_{(m)} \\ &= \arg\max_m \left\{ p(m) \int p(\mathbf{g}|\mathbf{f}_{(m)}, m)\, p(\mathbf{f}_{(m)}|m)\, d\mathbf{f}_{(m)} \right\} \\ &= \arg\max_m \{ p(m) \underbrace{p(\mathbf{g}|m)}_{\text{Evidence}} \} \end{aligned}$$

- Main difficulty: improper priors (for $p(\mathbf{f}_{(m)}|m)$) are not valid, because they are only defined up to a factor.

Bayesian model selection

- Comparing two models: which of m_1 or m_2 is *a posteriori* more likely?
- Answer is given by the so-called "posterior odds ratio"

$$\frac{p(m_1|\mathbf{g})}{p(m_2|\mathbf{g})} = \underbrace{\frac{p(\mathbf{g}|m_1)}{p(\mathbf{g}|m_2)}}_{\text{"Bayes' factor"}} \times \underbrace{\frac{p(m_1)}{p(m_2)}}_{\text{"prior odds ratio"}}$$

- Bayes' factor = evidence, provided by $\mathbf{g}$, for m_1 versus m_2.

Bayesian model selection: Example

Does a sequence of binary variables (e.g., coin tosses) comes from two different sources?

- Observations: $\mathbf{g} = [g_1, ..., g_t, g_{t+1}, ..., g2t]$, with $g_i \in \{0, 1\}$.
- Competing models:
  - m_1 ="all g_i's come from the same i.i.d. binary source with Prob(1) $= \alpha$" (e.g., same coin).
  - m_2 ="$[g_1, ..., g_t]$ and $[g_{t+1}, ..., g_{2t}]$ come from two different sources with Prob(1)= β and Prob(1) $= \gamma$, respectively"

    (e.g., two coins with different probabilities of "heads").
- Parameter vector under m_1, $\mathbf{f}_{(m_1)} = [\alpha]$

  Parameter vector under m_2, $\mathbf{f}_{(m_2)} = [\beta\ \gamma]$

  Notice that with $\beta = \gamma$, m_2 becomes m_1

Bayesian model selection: Example (cont.)

- Likelihood function under m_1:

$$p(\mathbf{g}|\alpha, m_1) = \prod_{i=1}^{2t} \alpha^{g_i}(1-\alpha)^{1-g_i} = \alpha^{n(\mathbf{g})}(1-\alpha)^{2t-n(\mathbf{g})}$$

  where $n(\mathbf{g})$ is the total number of 1's.

- Likelihood function under m_2:

$$p(\mathbf{g}|\beta, \gamma, m_2) = \beta^{n_1(\mathbf{g})}(1-\beta)^{t-n_1(\mathbf{g})}\ \gamma^{n_2(\mathbf{g})}(1-\gamma)^{t-n_2(\mathbf{g})}$$

  where $n_1(\mathbf{g})$ and $n_2(\mathbf{g})$ are the numbers of ones in the first and second halves of the data, respectively.

- Notice that $n_1(\mathbf{g}) + n_2(\mathbf{g}) = n(\mathbf{g})$.

Bayesian model selection: Example (cont.)

- Prior under m_1:

$$p(\alpha|m_1) = 1, \quad \text{for} \quad \alpha \in [0,1]$$

- Prior under m_2:

$$p(\beta, \gamma|m_2) = 1 \quad \text{for} \quad (\beta, \gamma) \in [0,1] \times [0,1]$$

- These two priors mean: "in any case, we know nothing about the parameters".

Bayesian model selection: Example (cont.)

- Evidence in favor of m_1 (recall that $p(\alpha|m_1) = 1$)

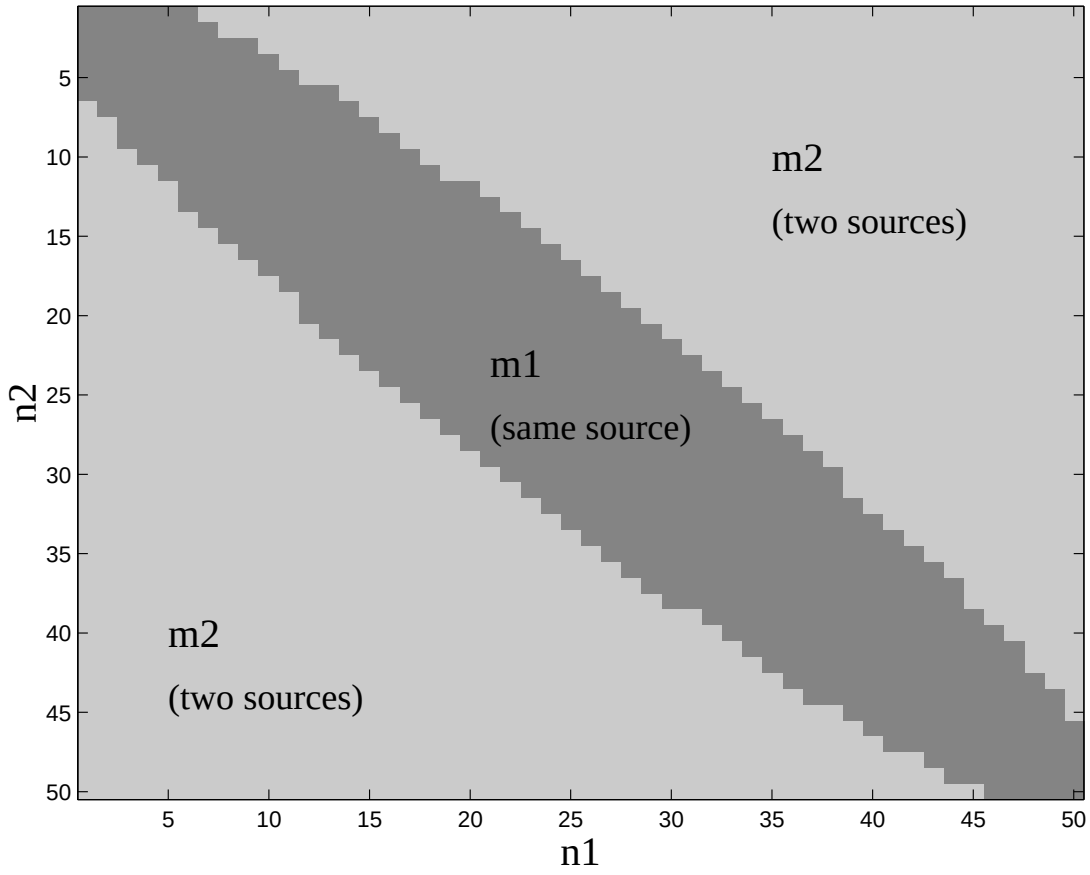
$$p(m_1|\mathbf{g}) = \int_0^1 \alpha^{n(\mathbf{g})}(1-\alpha)^{2t-n(\mathbf{g})}\, d\alpha = \frac{(2t-n(\mathbf{g}))!\, n(\mathbf{g})!}{(2a+1)!}$$

- Evidence in favor of m_2 (recall that $p(\beta, \gamma|m_2) = 1$):

$$\begin{aligned} p(m_2|\mathbf{g}) &= \int_0^1 \int_0^1 \beta^{n_1(\mathbf{g})}(1-\beta)^{t-n_1(\mathbf{g})}\ \gamma^{n_2(\mathbf{g})}(1-\gamma)^{t-n_2(\mathbf{g})}\, d\beta\, d\gamma \\ &= \frac{(t-n_1(\mathbf{g})!\, n_1(\mathbf{g})!}{(t+1)!}\, \frac{(t-n_2(\mathbf{g})!\, n_2(\mathbf{g})!}{(t+1)!} \end{aligned}$$

Bayesian model selection: Example (cont.)

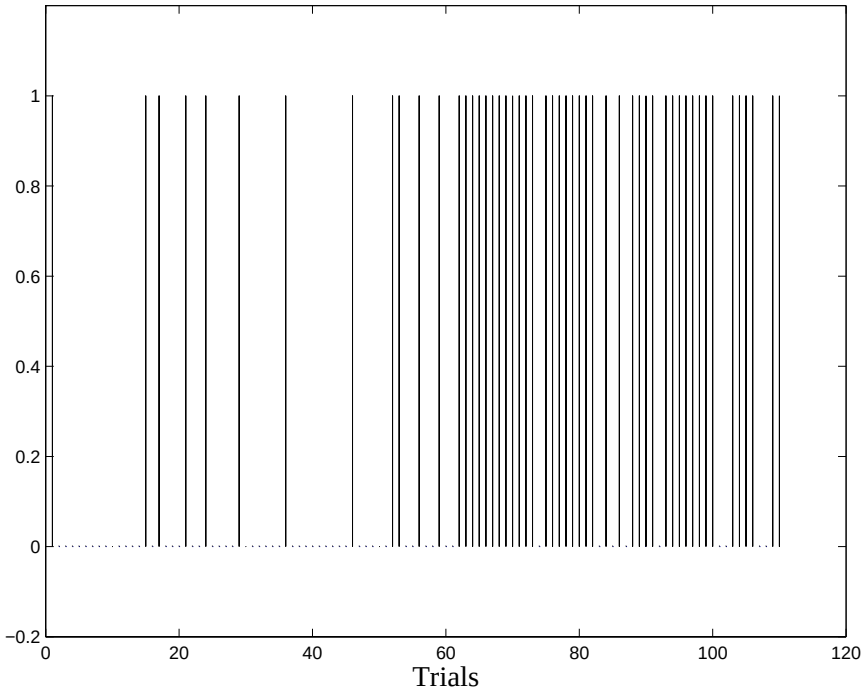
Decision regions for all possible outcomes with $2t = 100$, and $p(m_1) = p(m_2) = 1/2$.

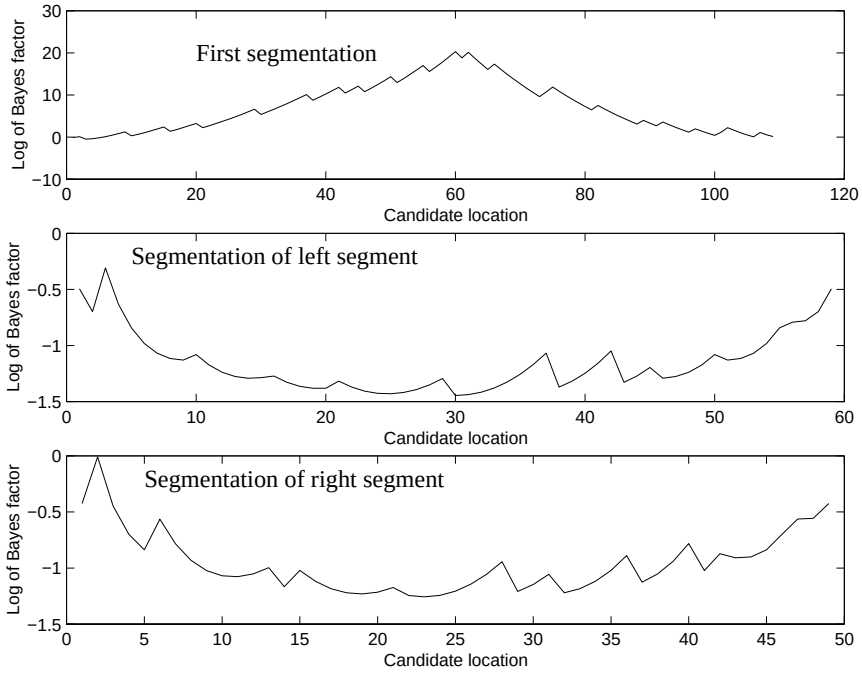


Bayesian model selection: Another example

Segmenting a sequence of binary i.i.d. observations:

Is there a change of model? Where?





Model selection: Schwarz's Bayesian inference criterion (BIC)

- Often, it is very difficult/impossible to compute $p(\mathbf{g}|m)$.
- By using a Taylor expansion of the likelihood, around the ML estimate, and for a "smooth enough" prior, we have

$$p(\mathbf{g}|m) \simeq p(\mathbf{g}|\widehat{\mathbf{f}}_{(m)}, m)\, n^{-\frac{\dim(\mathbf{f}_{(m)})}{2}} \equiv \text{BIC}(m)$$

  $\widehat{\mathbf{f}}_{(m)}$ is the ML estimate, under model m.

  $\dim(\mathbf{f}_{(m)})$ ="dimension of $\mathbf{f}_{(m)}$ under model m".

  n is the size of the observation vector $\mathbf{g}$.

- Let us also look at

$$-\log(\text{BIC}(m)) = -\log p(\mathbf{g}|\widehat{\mathbf{f}}_{(m)}, m) + \frac{\dim(\mathbf{f}_{(m)})}{2}\log n$$

Model selection: Rissanen's minimum description length (MDL)

- Consider an unknown $\mathbf{f}_{(k)}$ of unknown dimension k.
- Data is observed according to $p(\mathbf{g}|\mathbf{f}_{(k)})$
- For each k (each model), $p(\mathbf{f}_{(k)}|k)$ is constant;

  i.e., if k was known, we could find the ML estimate $\widehat{\mathbf{f}}_{(k)}$
- However, k is unknown, and the likelihood increases with k:

$$k_2 > k_1 \Rightarrow p(\mathbf{g}|\widehat{\mathbf{f}}_{(k_2)}) \geq p(\mathbf{g}|\widehat{\mathbf{f}}_{(k_1)})$$

- Conclusion: the ML estimate of k is: "as large as possible";

  this is clearly useless.

Minimum description length (MDL)

- Fact (from information theory): the shortest code-length for data $\mathbf{g}$ given that it was generated according to $p(\mathbf{g}|\mathbf{f}_{(k)})$ is

$$L(\mathbf{g}|\mathbf{f}_{(k)}) = -\log_2 p(\mathbf{g}|\mathbf{f}_{(k)}) \quad \text{(bits)}$$

- Then, for a given k, looking for the ML estimate of $\mathbf{f}_{(k)}$ is the same as looking for the code for which $\mathbf{g}$ has the shortest code-word:

$$\arg\max_{\mathbf{f}_{(k)}} p(\mathbf{g}|\mathbf{f}_{(k)}) = \arg\min_{\boldsymbol{\theta}_{(k)}} \left\{ -\log p(\mathbf{g}|\mathbf{f}_{(k)}) \right\} = \arg\min_{\mathbf{f}_{(k)}} L(\mathbf{g}|\mathbf{f}_{(k)})$$

- If a code is built to transmit $\mathbf{g}$, based on $\mathbf{f}_{(k)}$, then $\mathbf{f}_{(k)}$ also has to be transmitted. Conclusion: the total code-length is

$$L(\mathbf{g}, \mathbf{f}_{(k)}) = L(\mathbf{g}|\mathbf{f}_{(k)}) + L(\mathbf{f}_{(k)})$$

Minimum description length (MDL) (cont.)

- The total code-length is

$$L(\mathbf{g}, \mathbf{f}_{(k)}) = -\log_2 p(\mathbf{g}|\mathbf{f}_{(k)}) + L(\mathbf{f}_{(k)})$$

- The MDL criterion:

$$(\widehat{k}, \widehat{\mathbf{f}}_{(\widehat{k})})_{\text{MDL}} = \arg \min_{k, \mathbf{f}_{(k)}} \left\{ -\log_2 p(\mathbf{g}|\mathbf{f}_{(k)}) + L(\mathbf{f}_{(k)}) \right\}$$

- Basically, the term $L(\mathbf{f}_{(k)})$ grows with k counterbalancing the behavior of the likelihood.

- From a Bayesian point of view, we have a prior

$$p(\mathbf{f}_{(k)}) \propto 2^{-L(\mathbf{f}_{(k)})}$$

Minimum description length (cont.)

- What about $L(\mathbf{f}_{(k)})$? It is problem-dependent.
- If the components of $\mathbf{f}_{(k)}$ are real numbers (and under certain other conditions) the (asymptotically) optimal choice is

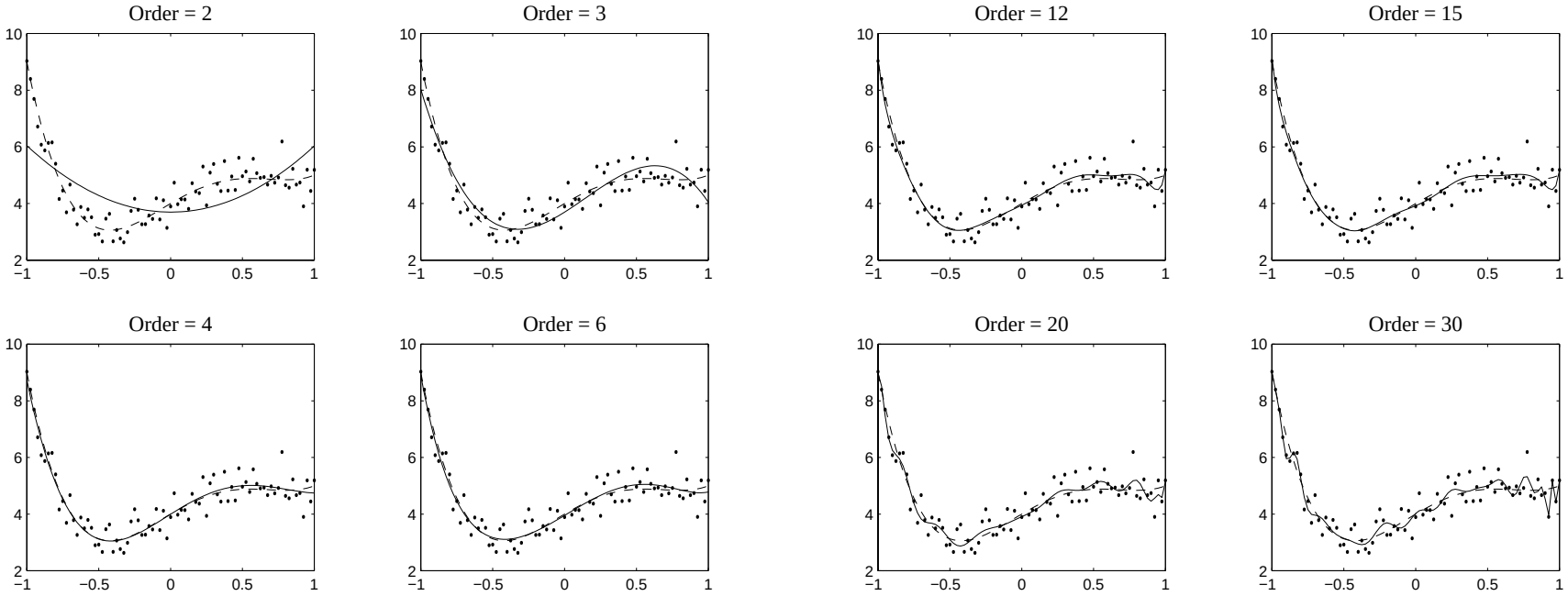
$$L(\mathbf{f}_{(k)}) = \frac{k}{2} \log n$$

  where n is the size of the data vector $\mathbf{g}$.
- Interestingly, in this case MDL coincides with BIC.
- In other situations (e.g., discrete parameters), there are natural choices.

Minimum description length: Example

Fitting a polynomial of unknown degree: $\mathbf{f}_{(k+1)}$ contains the coefficients of a k-order polynomial.

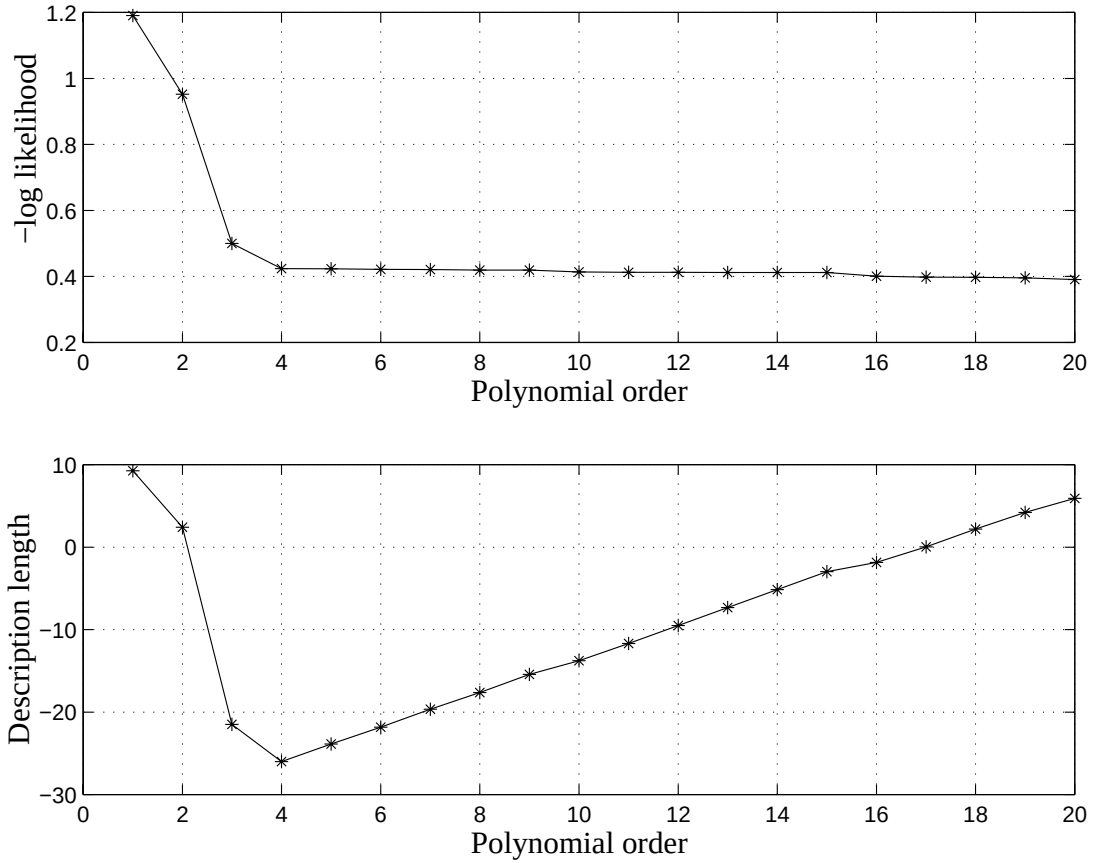
Observation model: $\mathbf{g}$ = "true polynomial plus white Gaussian noise".



Minimum description length: Example

Fitting a polynomial of unknown degree.

$-\log p(\mathbf{g}|\mathbf{f}_{(k)})$ keeps going down, but MDL picks the right order $\widehat{k} = 4$.



Introduction to Markov Random Fields

- Image analysis problems $\Leftrightarrow$ compound inference problems.
- Prior $p(\mathbf{f})$ formalizes expected joint behavior of elements of $\mathbf{f}$.
- Markov random fields: a convenient tool to write priors for image analysis problems.
- Just as Markov random processes formalize temporal evolutions/dependencies.

Graphs and random fields on graphs.

Basic graph-theoretic concepts

- A *graph* $\mathbf{G} = (\mathbf{N}, \mathbf{E})$ is a collection of *nodes* (or *vertices*) $\mathbf{N} = \{n_1, n_2, ...n_{|\mathbf{N}|}\}$

  and *edges* $\mathbf{E} = \{(n_{i_1}, n_{i_2}), ...(n_{i_{2|\mathbf{E}|-1}}, n_{i_{2|\mathbf{E}|}})\} \subseteq \mathbf{N} \times \mathbf{N}$.

  Notation: $|\mathbf{N}|$ = number of elements of set $\mathbf{N}$.

- We consider only undirected graphs, i.e., the elements of $\mathbf{E}$ are seen as unordered pairs: $(n_i, n_j) \equiv (n_j, n_i)$.

- Two nodes n_1, $n_2 \in \mathbf{N}$ are *neighbors* if the corresponding edge exists, i.e., if $(n_1, n_2) \in \mathbf{E}$.

Graphs and random fields on graphs.

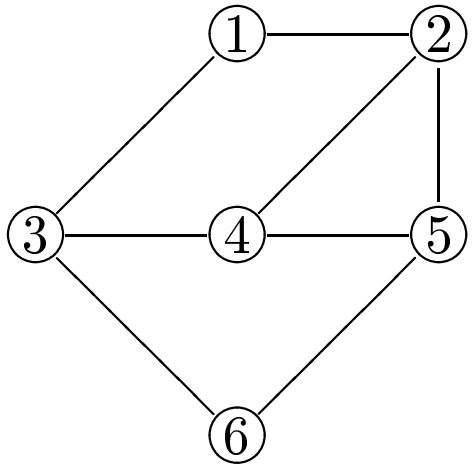
Basic graph-theoretic concepts (cont.)

- A complete graph: all nodes are neighbors of all other nodes.
- A node is not neighbor of itself; no (n_i, n_i) edges are allowed.
- Neighborhood of a node: $N(n_i) = \{n_j : (n_i, n_j) \in \mathbf{E}\}$.
- The neighborhood relation is symmetrical:

$$n_j \in N(n_i) \Leftrightarrow n_i \in N(n_j)$$

Graphs and random fields on graphs.

Example of a graph:



$\mathbf{N} = \{1, 2, 3, 4, 5, 6\}$

$\mathbf{E} = \{(1,2), (1,3), (2,4), (2,5), (3,6), (5,6), (3,4), (4,5)\} \subset \mathbf{N} \times \mathbf{N}$

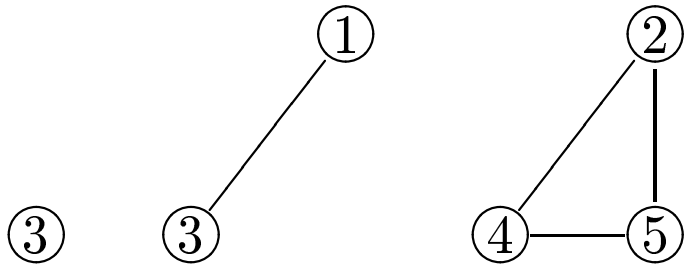
$N(1) = \{2, 3\}$, $N(2) = \{1, 4, 5\}$, $N(3) = \{1, 4, 6\}$, etc...

Graphs and random fields on graphs.

- *Clique* of **G** is either a single node or a complete subgraph of **G**.

  In other words, a single node or a subset of nodes that are all mutual neighbors.

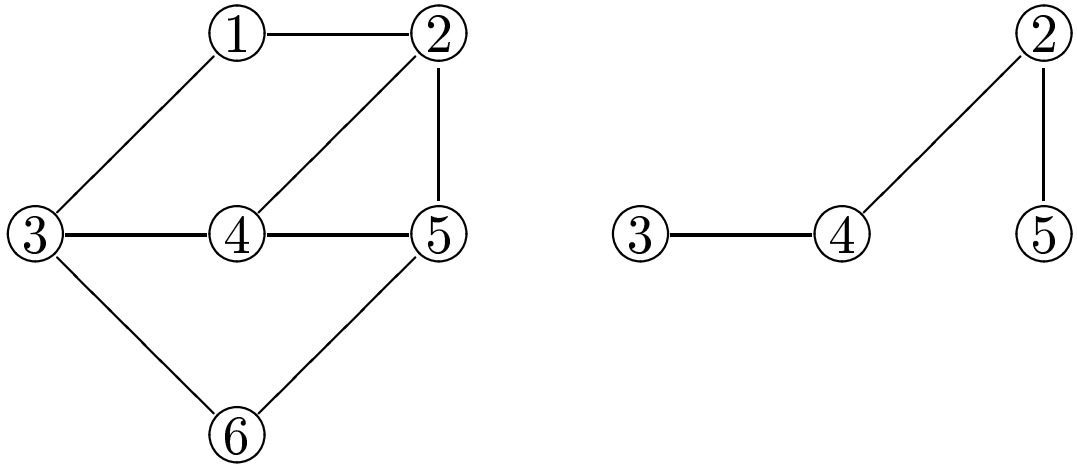
- Examples of cliques from the previous graph



- Set of all cliques (from the same example): $\mathcal{C} = \mathbf{N} \cup \mathbf{E} \cup \{(2, 4, 5)\}$

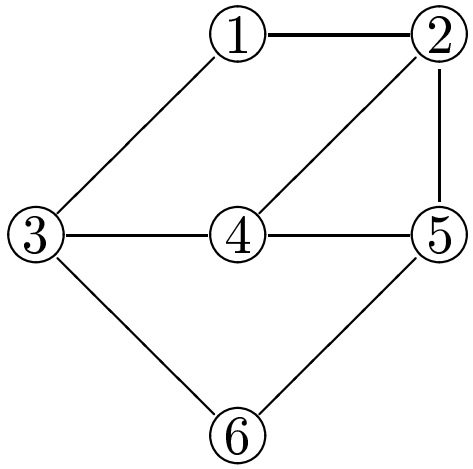
Graphs and random fields on graphs.

- A length-k *path* in $\mathbf{G}$ is an ordered sequence of nodes, $(n_1, n_2, ...n_k)$, such that $(n_j, n_{j+1}) \in \mathbf{E}$.
- Example: a graph and a length-4 path.



Graphs and random fields on graphs.

- Let **A**, **B**, **C** be three disjoint subsets of **N**.
- We say that **C** *separates* **A** from **B** if any path from a node in **A** to a node in **B** contains one (or more) node in **C**.
- Example, in the graph



$\mathbf{C} = \{1, 4, 6\}$ separates $\mathbf{A} = \{3\}$ from $\mathbf{B} = \{2, 5\}$

Graphs and random fields on graphs.

- Consider a joint probability function $p(\mathbf{f}) = p(f_1, f_2, ..., f_m)$..
- Assign each variable to a node of a graph, $\mathbf{N} = \{1, 2, ..., m\}$.
  We have "random field on graph $\mathbf{N}$".
- Let $\mathbf{f_A}$, $\mathbf{f_B}$, $\mathbf{f_C}$ be three disjoint subsets of $\mathbf{F}$ (i.e., $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are disjoint subsets of $\mathbf{N}$. If

  $p(\mathbf{f_A}, \mathbf{f_B}|\mathbf{f_C}) = p(\mathbf{f_A}|\mathbf{f_C})p(\mathbf{f_B}|\mathbf{f_C}) \quad \Leftarrow \quad$ "$\mathbf{C}$ separates $\mathbf{A}$ from $\mathbf{B}$".

  "$p()$ is global Markov" with respect to $\mathbf{N}$. The graph is called an "I-map" of $p(\mathbf{f})$
- Any $p(\mathbf{f})$ is "global Markov" with respect to the complete graph.
- If rather than $\Leftarrow$, we have $\Leftrightarrow$, the graph is called a "perfect I-map".

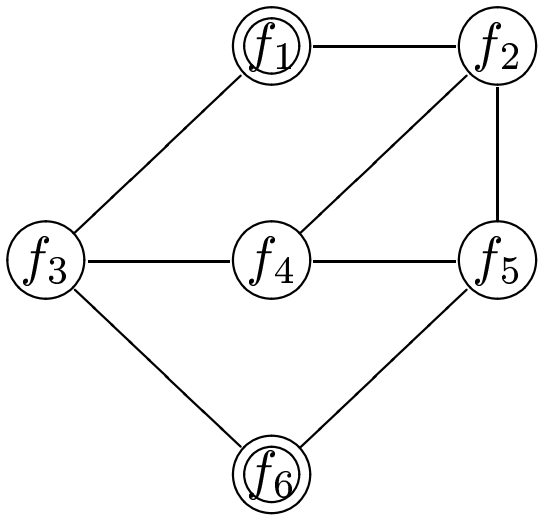
Graphs and random fields on graphs.

Pair-wise Markovianity.

- **Pair-wise Markovianity:** $(i,j) \notin \mathbf{E} \Rightarrow$ "f_i and f_j are independent, when conditioned on all the other variables".

  Proof: simply notice that if i and j are not neighbors, the remaining nodes *separate* i from j.

  Example: in the following graph,
  $p(f_1, f_6|f_2, f_3, f_4, f_5) = p(f_1|f_2, f_3, f_4, f_5)p(f_6|f_2, f_3, f_4, f_5)$.



Local Markovianity.

- **Local Markovianity:**
  $p(f_i, f_{\mathbf{N}/(\{i\}\cup N(i))}|f_{N(i)}) = p(f_i|f_{N(i)})\, p(f_{\mathbf{N}/(\{i\}\cup N(i))}|f_{N(i)})$;
  "given its neighborhood, a variable is independent on the rest".

  Proof: Notice that $N(f_i)$ separates f_i from the rest of the graph.

- Equivalent form (better known in the MRF literature):

$$p(f_i|f_{\mathbf{N}/\{i\}}) = p(f_i|f_{N(i)})$$

  Proof: divide the above equality by $p(f_{\mathbf{N}/(\{i\}\cup N(i))}|f_{N(i)})$:

$$\begin{aligned} \frac{p(f_i, f_{\mathbf{N}/(\{i\}\cup N(i))}|f_{N(i)})}{p(f_{\mathbf{N}/(\{i\}\cup N(i))}|f_{N(i)})} &= p(f_i|f_{N(i)}) \\ p(f_i|f_{\mathbf{N}/\{i\}}) &= p(f_i|f_{N(i)}) \end{aligned}$$

  because $[\mathbf{N}/(\{i\}\cup N(i))]\cup N(i) = \mathbf{N}/\{i\}$.

Hammersley-Cliford theorem

Consider a random field $\mathbf{F}$ on a graph $\mathbf{N}$, such that $p(\mathbf{f}) > 0$.

a) If the field $\mathbf{F}$ has the local Markov property, then $p(\mathbf{f})$ can be written as a Gibbs distribution

$$p(\mathbf{f}) = \frac{1}{Z} \exp \left\{ - \sum_{C \in \mathcal{C}} V_C(\mathbf{f}_C) \right\}$$

where Z, the normalizing constant, is called the *partition function*. The functions $V_C(\cdot)$ are called *clique potentials*. The negative of the exponent is called *energy*.

b) If $p(\mathbf{f})$ can be written in Gibbs form for the cliques of some graph, then it has the global Markov property.

Fundamental consequence: a Markov random field can be specified via the clique potentials.

Hammersley-Cliford theorem (cont.)

- Computing the local Markovian conditionals from the clique potentials

$$p(f_i|f_{N(i)}) = \frac{1}{Z(f_{N(i)})} \exp\left\{ - \sum_{C:i\in C} V_C(\mathbf{f}_C) \right\}$$

- Notice that the normalizing constant may depend on the neighborhood state.

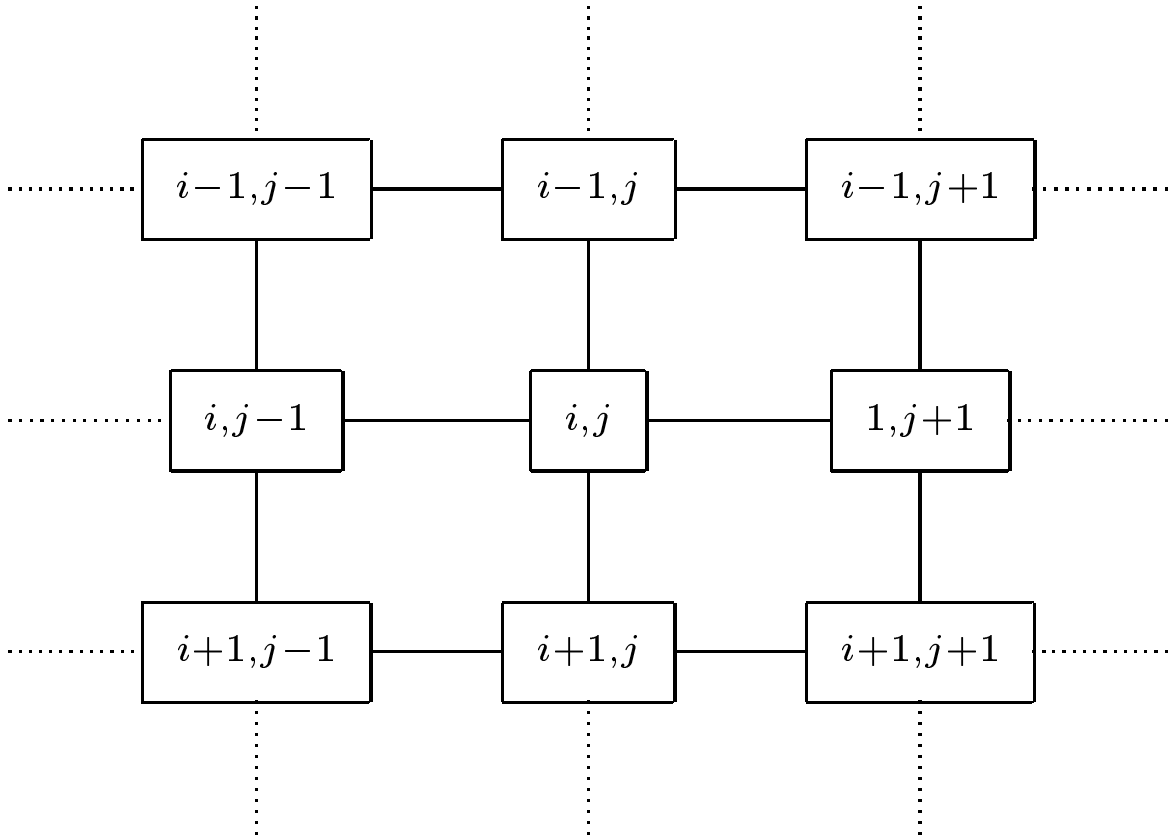
Regular rectangular lattices

- Let us now focus on *regular rectangular lattices*.
  $\mathbf{N} = \{(i,j),\ \ i = 1, ..., M,\ j = 1, ..., N\}$

- A hierarchy neighborhood systems:
  $N^0(i,j) = \{\ \}$, zero-order (empty neighborhoods);
  $N^1(i,j) = \{(k,l),\ \ (i-k)^2 + (j-l)^2 \leq 1\}$, order-1 (4 nearest neighbors);
  $N^2(i,j) = \{(k,l),\ \ (i-k)^2 + (j-l)^2 \leq 2\}$, order-2 (8 nearest neighbors);
  etc...

Regular rectangular lattices

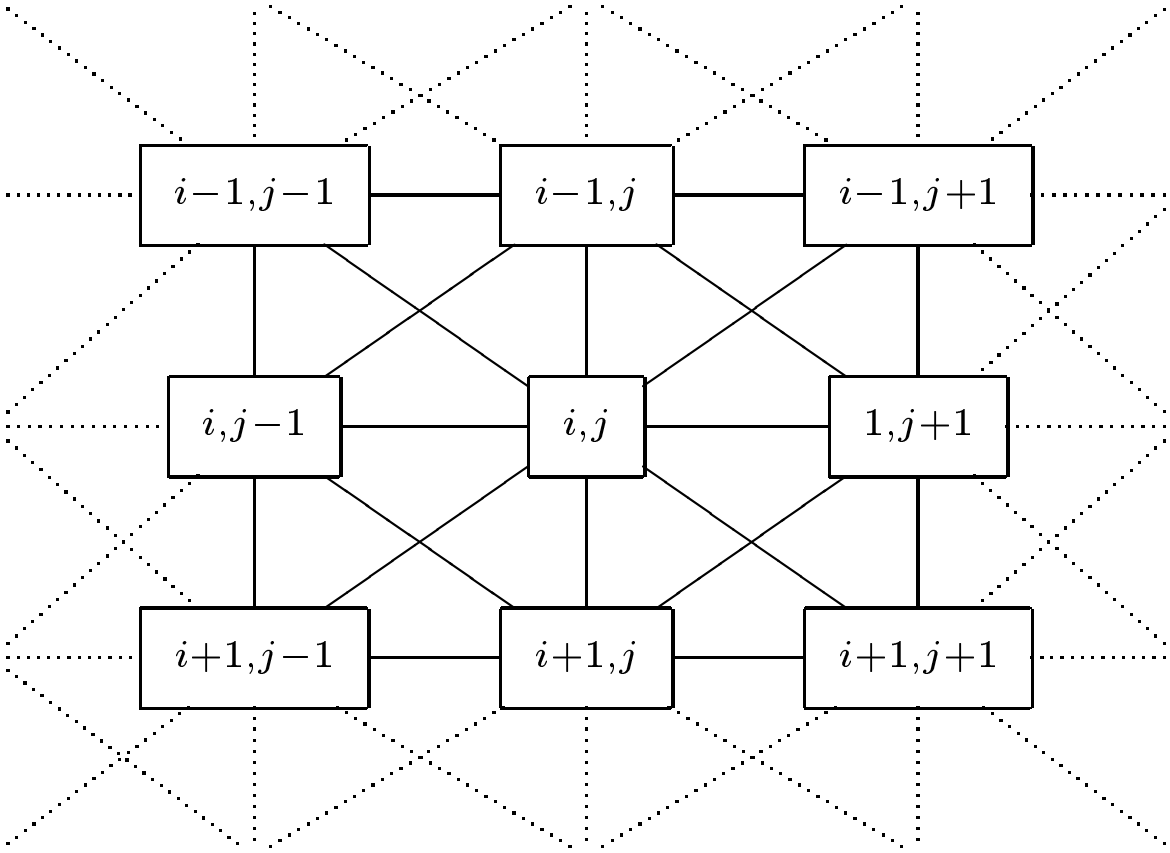
Illustration of first order neighborhood system:



$N^1(i,j) = \{(i-1,j),(i,j-1),(i,j+1),(i+1,j)\}$ (4 nearest neighbors).

Regular rectangular lattices

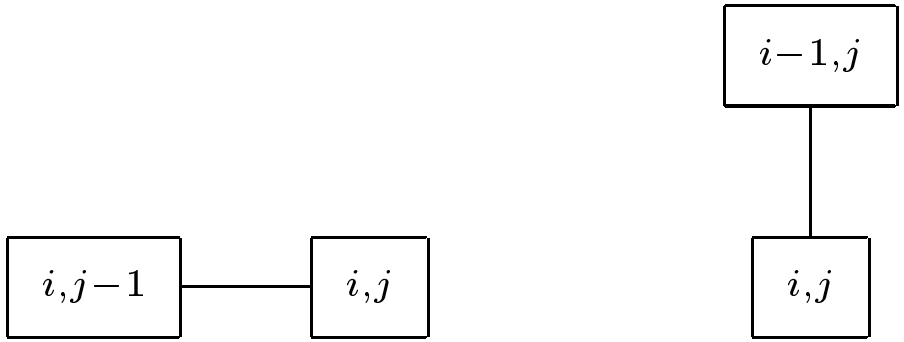
Illustration of second order neighborhood system:



$N^2(i,j) = \{(i-1,j-1),(i-1,j),(i-1,j+1),(i,j-1),(i,j+1),(i+1,j-1),(i+1,j),(i+1,j+1)\}$
(8 nearest neighbors).

Regular rectangular lattices

Cliques of a first order neighborhood system: all single nodes plus all subgraphs of the types

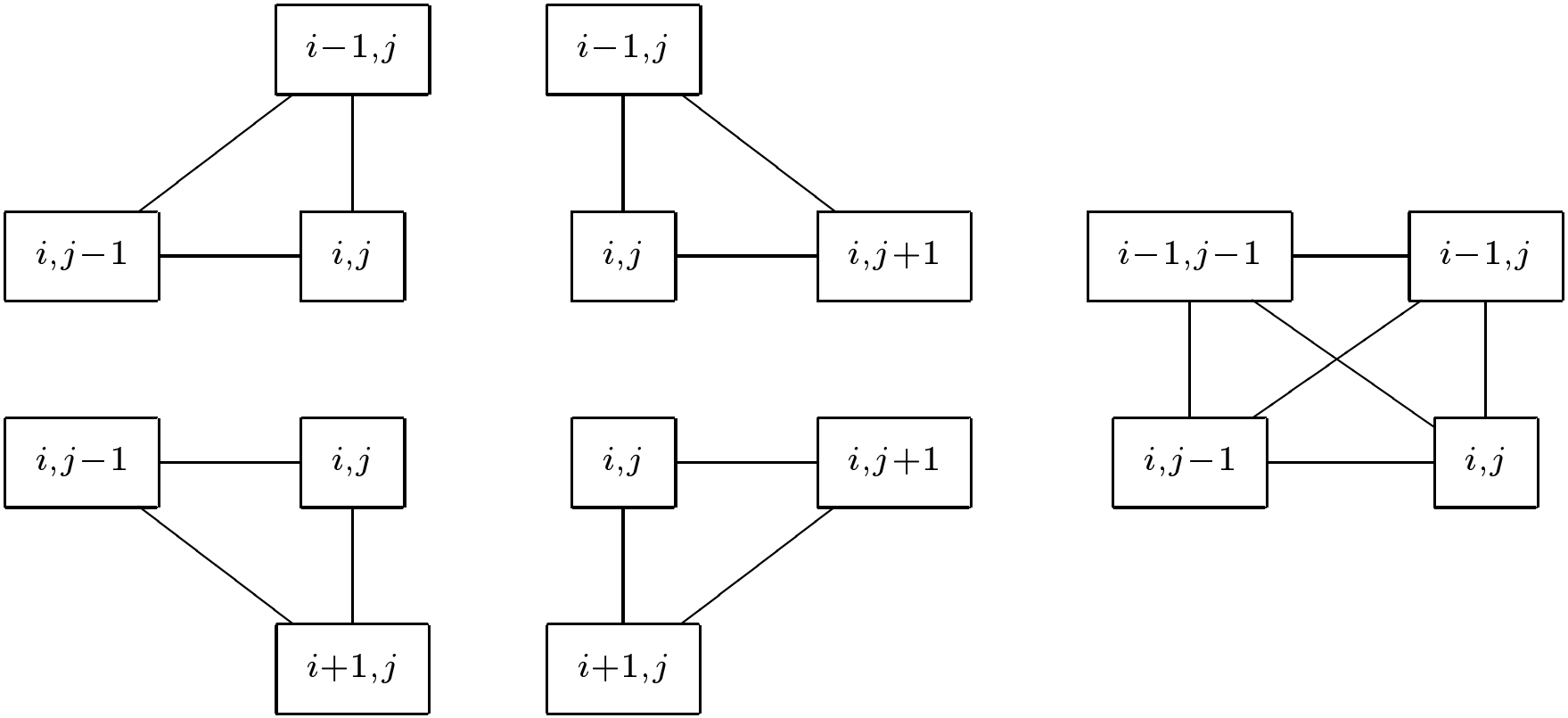


Notation:

$\mathcal{C}^k$ = "set of all cliques for the order-k neighborhood system".

Regular rectangular lattices

Cliques of a second order neighborhood: $\mathcal{C}^1$ plus all subgraphs of the types



Auto-models

- Only pair-wise interactions.
- In terms of clique potentials: $|C| > 2 \Rightarrow V_C(\cdot) = 0$.
- These are the simples models, beyond site independence.
- Even for large neighborhoods, we can define an auto-model.

Gauss-Markov Random Fields (GMRF)

- Joint probability density function (for zero mean)

$$p(\mathbf{f}) = \frac{\sqrt{\det(\mathbf{A})}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\mathbf{f}^T\mathbf{A}\mathbf{f}\right\}$$

- The quadratic form in the exponent can be written as

$$\mathbf{f}^T\mathbf{A}\mathbf{f} = \sum_{i=1}^{m}\sum_{j=1}^{m} f_i f_j A_{ij}$$

  revealing that this is an auto-model (there are only pair-wise terms).

- Matrix $\mathbf{A}$ (the *potential* matrix, inverse of the covariance matrix) determines the neighborhood system:

$$i \in N(j) \Leftrightarrow A_{ij} \neq 0$$

Notice that to be a valid potential matrix, $\mathbf{A}$ has to be symmetric, thus respecting the symmetry of neighborhood relations.

Gauss-Markov Random Fields (GMRF)

- Local (Markov-type) conditionals are univariate Gaussian

$$
\begin{aligned}
p(f_i|\{f_j; j \neq i\}) &= \sqrt{\frac{A_{ii}}{2\pi}} \exp\left\{ -\frac{A_{ii}}{2} \left(f_i - \frac{1}{A_{ii}} \sum_{j \neq i} A_{ij} f_j \right)^2 \right\} \\
&\sim \mathcal{N}\left(\frac{1}{A_{ii}} \sum_{j \neq i} A_{ij} f_j, \frac{1}{A_{ii}} \right)
\end{aligned}
$$

Gauss-Markov Random Fields (GMRF)

- Specification via clique-potentials: squares of differences,

$$V_C(\mathbf{f}_C) = \frac{\mu}{2}(\sum_{j \in C} \alpha_j^C f_j)^2 = \frac{\mu}{2}(\sum_{j \in \mathbf{N}} \alpha_j^C f_j)^2$$

  as long as we define $\alpha_j^C = 0 \Leftarrow j \notin C$.

- The exponent of the GMRF density becomes

$$\begin{aligned} -\sum_{C \in \mathcal{C}} V_C(\mathbf{f}) &= -\frac{\mu}{2} \sum_{C \in \mathcal{C}} \left(\sum_{j \in \mathbf{N}} \alpha_j^C f_j \right)^2 \\ &= -\frac{\mu}{2} \sum_{j \in \mathbf{N}} \sum_{k \in \mathbf{N}} \underbrace{\left(\sum_{C \in \mathcal{C}} \alpha_j^C \alpha_k^C \right)}_{A_{j\,k}} f_j f_k \equiv -\frac{\mu}{2} \mathbf{f}^T \mathbf{A} \mathbf{f} \end{aligned}$$

  showing this is a GMRF with potential matrix $\frac{\mu}{2}\mathbf{A}$

Gauss-Markov Random Fields (GMRF): The classical "smoothing prior" GMRF.

- A lattice $\mathbf{N} = \{(i,j),\ \ i = 1, ..., M,\ j = 1, ..., N\}$
- A first order neighborhood
  $N((i,j)) = \{(i-1,j), (i,j-1), (i+1,j), (i,j+1)\}$
- Clique set: all pairs of (vertically or horizontally) adjacent sites.
- Clique-potentials: squares of first-order differences,

$$V_{\{(i,j),(i,j-1)\}}(f_{i\,j}, f_{i\,j-1}) = \frac{\mu}{2}(f_{i\,j} - f_{i\,j-1})^2$$

$$V_{\{(i,j),(i-1,j)\}}(f_{i\,j}, f_{i-1\,j}) = \frac{\mu}{2}(f_{i\,j} - f_{i-1\,j})^2$$

- Resulting $\mathbf{A}$ matrix: block-tridiagonal with tridiagonal blocks.
- Matrix $\mathbf{A}$ is also quasi-block-Toeplitz with quasi-Toeplitz blocks. "Quasi-" due to boundary corrections.

Bayesian image restoration with GMRF prior:

- A "smoothing" GMRF prior: $p(\mathbf{f}) \propto \exp\left\{-\frac{\mu}{2}\mathbf{f}^T\mathbf{A}\mathbf{f}\right\}$ where $\mathbf{A}$ is as defined in the previous slide.
- Observation model: linear operator (matrix) plus additive white Gaussian noise,

$$\mathbf{g} = \mathbf{H}\,\mathbf{f} + \mathbf{n}, \qquad \text{where} \quad \mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I}).$$

  Models well: out-of-focus blur, motion blur, tomographic imaging, ...
- There is nothing new: we saw before that the MAP and PM estimates are simply:

$$\widehat{\mathbf{f}} = \left[\mu\sigma^2\mathbf{A} + \mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{g}$$

  ...only difficulty: the matrix to be inverted is huge.

Bayesian image restoration with GMRF prior (cont.)

- With a "smoothing" GMRF prior and a linear observation model plus Gaussian noise, optimal estimate:

$$\widehat{\mathbf{f}} = \left[\mu\sigma^2\mathbf{A} + \mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{g}$$

- A similar result can be obtained in other theoretical frameworks: regularization, penalized-likelihood.

- Notice that

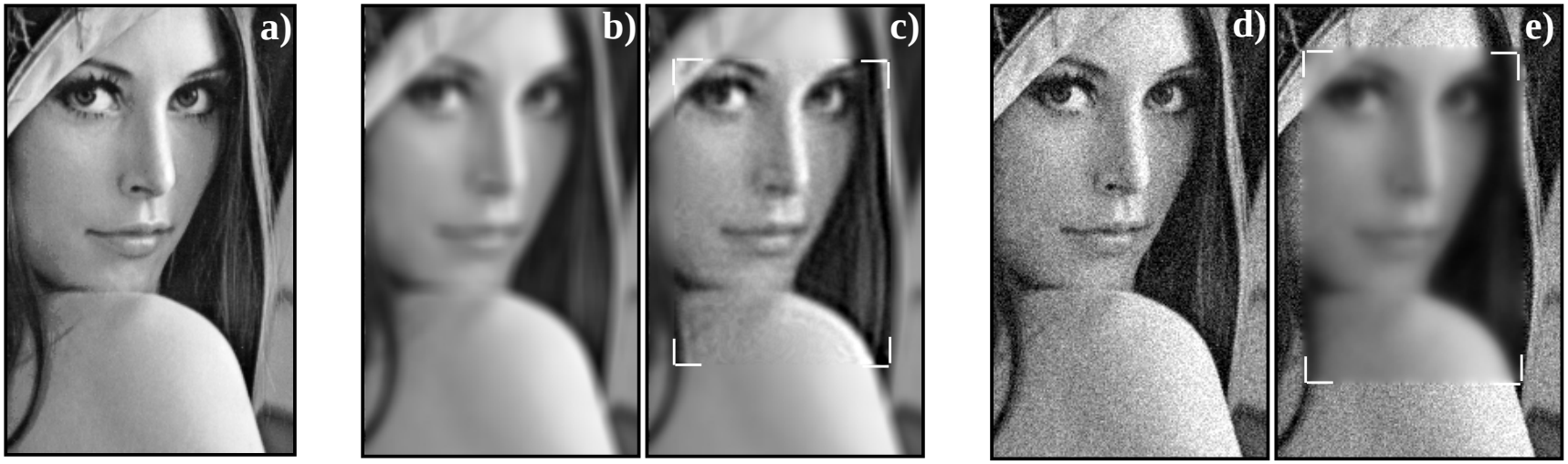
$$\lim_{\mu\to 0}\left[\mu\sigma^2\mathbf{A} + \mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T = \left[\mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T \equiv \mathbf{H}^\dagger$$

  the (least squares) pseudo-inverse of $\mathbf{H}$.

- The huge size of $\left[\mu\sigma^2\mathbf{A} + \mathbf{H}^T\mathbf{H}\right]$ precludes any explicit inversion. Iterative schemes are (almost always) used.

Bayesian image restoration with GMRF prior (cont.)

Examples:



(a) Original; (b) blurred and slightly noisy; (c) restored from (b); (d) no blur, severe noise; (e) restored from (d).

Deblurring: good job. Denoising: oversmoothing, i.e. "discontinuites" are smoothed out.

Solutions to the oversmoothing nature of the GMRF prior.

- Explicitly detect and preserve discontinuities: *compound GMRF* models, weak-membrane, etc...

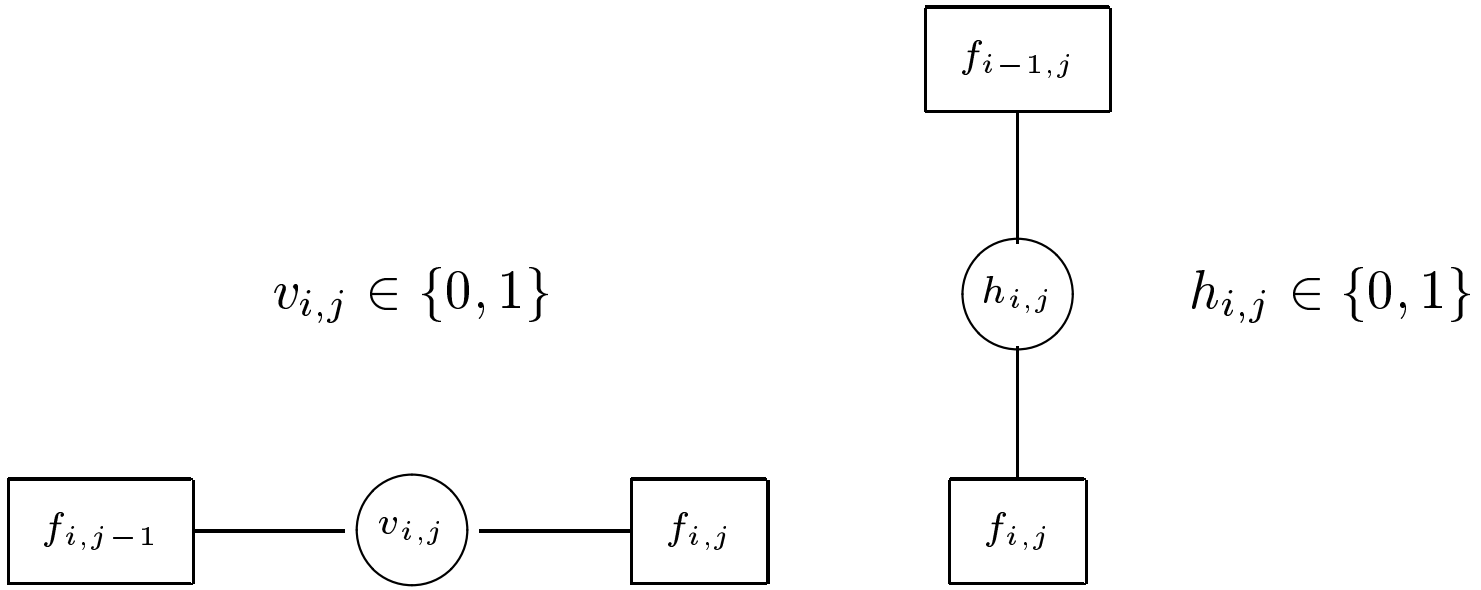
  A new set of variables comes into play: the edge (or line) field.

- Replace the "square law" potentials by other more "robust" functions.

  The quadratic nature of the *a posteriori* energy is lost.
  Consequence: optimization becomes much more difficult.

Compound Gauss-Markov random fields

- Insert a binary variables which can "turn off" clique potentials.



- New clique potentials:

$$V(f_{i\,j}, f_{i\,j-1}, v_{i\,j}) = \frac{\mu}{2}(1 - v_{i\,j})(f_{i\,j} - f_{i\,j-1})^2$$

$$V(f_{i\,j}, f_{i-1\,j}, h_{i\,j}) = \frac{\mu}{2}(1 - h_{i\,j})(f_{i\,j} - f_{i-1\,j})^2$$

Compound Gauss-Markov random fields (cont.)

- The line variables can "turn on" the quadratic potentials,

$$\begin{aligned} V(f_{i\,j}, f_{i\,j-1}, 0) &= \frac{\mu}{2}(f_{i\,j} - f_{i\,j-1})^2 \\ V(f_{i\,j}, f_{i-1\,j}, 0) &= \frac{\mu}{2}(f_{i\,j} - f_{i-1\,j})^2 \end{aligned}$$

  or "turn them off",

$$\begin{aligned} V(f_{i\,j}, f_{i\,j-1}, 1) &= 0 \\ V(f_{i\,j}, f_{i-1\,j}, 1) &= 0 \end{aligned}$$

  meaning, "there is an edge here, do not smooth!".

Compound Gauss-Markov random fields (cont.)

- Given a certain configuration of line variables, we still have a Gauss Markov random field

$$p(\mathbf{f}|\mathbf{h},\mathbf{v}) \propto \exp\left\{-\frac{\mu}{2}\mathbf{f}^T\mathbf{A}(\mathbf{h},\mathbf{v})\mathbf{f}\right\}$$

  but the potential matrix now depends on $\mathbf{h}$ and $\mathbf{v}$.

- Given $\mathbf{h}$ and $\mathbf{v}$, the MAP (and PM) estimate of $\mathbf{f}$ has the same form:

$$\widehat{\mathbf{f}}(\mathbf{h},\mathbf{v}) = \left[\mu\sigma^2\mathbf{A}(\mathbf{h},\mathbf{v}) + \mathbf{H}^T\mathbf{H}\right]^{-1}\mathbf{H}^T\mathbf{g}$$

- Question: how to estimate $\mathbf{h}$ and $\mathbf{v}$?

  Hint: $\mathbf{h}$ and $\mathbf{v}$ are "parameters" of the prior.

  This motivates a detour on: "how to estimate parameters?"

Parameter estimation in Bayesian inference problems

- The likelihood (observation model) depends on parameter(s) $\boldsymbol{\phi}$, i.e., we write $p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})$.
- The prior depends on parameter(s) $\boldsymbol{\psi}$, i.e., we write $p(\mathbf{f}|\boldsymbol{\psi})$.
- With explicit reference to these parameters, Bayes rule becomes:

$$p(\mathbf{f}|\mathbf{g}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \frac{p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})\ p(\mathbf{f}|\boldsymbol{\psi})}{\int p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})\ p(\mathbf{f}|\boldsymbol{\psi}) d\mathbf{f}} = \frac{p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})\ p(\mathbf{f}|\boldsymbol{\psi})}{p(\mathbf{g}|\boldsymbol{\phi}, \boldsymbol{\psi})}$$

- Question: how can we estimate $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ from $\mathbf{g}$, without violating the fundamental "likelihood principle"?

Parameter estimation in Bayesian inference problems

- How to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ from $\mathbf{g}$, without violating the "likelihood principle"?
- Answer: the scenario has to be modified.
  - Rather than just $\mathbf{f}$ there is a new set of unknowns: $(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})$.
  - There is a new likelihood function: $p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}) = p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})$.
  - A new prior is needed: $p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}) = p(\mathbf{f}|\boldsymbol{\psi})\, p(\boldsymbol{\phi}, \boldsymbol{\psi})$, because $\mathbf{f}$ is independent of $\boldsymbol{\phi}$.
    Usually, $p(\boldsymbol{\phi}, \boldsymbol{\psi})$ is called a hyper-prior.
- This is called a hierarchical Bayesian setting; here, with two levels.
  To add one more level, consider parameters $\boldsymbol{\alpha}$ of the hyper-prior $p(\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\alpha}) = p(\boldsymbol{\phi}, \boldsymbol{\psi}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha})$. And so on...
- Usually, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are *a priori* independent, $p(\boldsymbol{\phi}, \boldsymbol{\psi}) = p(\boldsymbol{\phi})\, p(\boldsymbol{\psi})$.

Parameter estimation in Bayesian inference problems

- We may compute a complete *a posterior* probability function:

$$
\begin{aligned}
p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{g}) &= \frac{p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})\, p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})}{\int\int\int p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})\, p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}) d\mathbf{f} d\boldsymbol{\phi} d\boldsymbol{\psi}} \\
&= \frac{p(\mathbf{g}|\mathbf{f}, \boldsymbol{\phi})\, p(\mathbf{f}|\boldsymbol{\psi})\, p(\boldsymbol{\phi}, \boldsymbol{\psi})}{p(\mathbf{g})}
\end{aligned}
$$

- How to use it, depends on the adopted loss function.
- Notice that, even if $\mathbf{f}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ are scalar, this is now a compound inference problem.

Parameter estimation in Bayesian inference problems

Non-additive "0/1" loss function $L\left[(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi}), (\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})\right]$.

- As seen above, this leads to the *joint* MAP (JMAP) criterion:

$$(\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{JMAP}} = \arg \max_{(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi} | \mathbf{g})$$

- With a uniform prior on the parameters $p(\boldsymbol{\phi}, \boldsymbol{\psi}) = k$,

$$\begin{aligned}
(\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{JMAP}} &= \arg \max_{(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi} | \mathbf{g}) \\
&= \arg \max_{(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{g} | \mathbf{f}, \boldsymbol{\phi})\, p(\mathbf{f} | \boldsymbol{\psi}) \\
&= \arg \max_{(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{g}, \mathbf{f} | \boldsymbol{\phi}, \boldsymbol{\psi}) \equiv (\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{GML}}
\end{aligned}$$

sometimes called the *generalized maximum likelihood* (GML).

Parameter estimation in Bayesian inference problems

A "0/1" loss function, additive with respect to $\mathbf{f}$ and the parameters, i.e.,

$$L\left[(\mathbf{f},\boldsymbol{\phi},\boldsymbol{\psi}),(\widehat{\mathbf{f}},\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\psi}})\right] = L^{(1)}\left[\mathbf{f},\widehat{\mathbf{f}}\right] + L^{(2)}\left[(\boldsymbol{\phi},\boldsymbol{\psi}),(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\psi}})\right]$$

$L^{(1)}[\cdot,\cdot]$ is a non-additive "0/1" loss function;

$L^{(2)}[\cdot,\cdot]$ is an arbitrary loss function.

- From the results above on additive loss functions, the estimate of $\mathbf{f}$ is

$$\begin{aligned}\widehat{\mathbf{f}}_{\text{MMAP}} &= \arg\max_{\mathbf{f}} \int\int p(\mathbf{f},\boldsymbol{\phi},\boldsymbol{\psi}|\mathbf{g})\, d\boldsymbol{\phi}\, d\boldsymbol{\psi} \\ &= \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{g})\end{aligned}$$

  the so-called *marginalized* MAP (MMAP).

- The parameters are "integrated out" from the *a posteriori* density.

Parameter estimation in Bayesian inference problems

As in the previous case, let

$$L\left[(\mathbf{f},\boldsymbol{\phi},\boldsymbol{\psi}),(\widehat{\mathbf{f}},\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\psi}})\right] = L^{(1)}\left[\mathbf{f},\widehat{\mathbf{f}}\right] + L^{(2)}\left[(\boldsymbol{\phi},\boldsymbol{\psi}),(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\psi}})\right]$$

now, with $L^{(2)}[\cdot,\cdot]$ a non-additive "0/1" loss function.

- Considering a uniform prior $p(\boldsymbol{\phi},\boldsymbol{\psi}) = k$,

$$\begin{aligned}(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\psi}})_{\text{MMAP}} &= \arg\max_{(\boldsymbol{\phi},\boldsymbol{\psi})} \int p(\mathbf{f},\boldsymbol{\phi},\boldsymbol{\psi}|\mathbf{g})\, d\mathbf{f} \\ &= \arg\max_{(\boldsymbol{\phi},\boldsymbol{\psi})} \int p(\mathbf{g}|\mathbf{f},\boldsymbol{\phi})\, p(\mathbf{f}|\boldsymbol{\psi})\, d\mathbf{f} \\ &= \arg\max_{(\boldsymbol{\phi},\boldsymbol{\psi})} \int p(\mathbf{g},\mathbf{f}|\boldsymbol{\phi},\boldsymbol{\psi})\, d\mathbf{f} \;=\; \arg\max_{(\boldsymbol{\phi},\boldsymbol{\psi})} p(\mathbf{g}|\boldsymbol{\phi},\boldsymbol{\psi})\end{aligned}$$

  the so-called *marginal maximum likelihood* (MML) estimate.

- The unknown image is "integrated out" from the likelihood function.

Parameter estimation in Bayesian inference problems

Implementing JMAP : $(\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{JMAP}} = \arg \max_{(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{f}, \boldsymbol{\phi}, \boldsymbol{\psi} | \mathbf{g})$

- This is usually very difficult to implement.
- A sub-optimal criterion, called *partial optimal solution* (POS):

$$(\widehat{\mathbf{f}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{POS}} = \text{solution of} \begin{cases} \widehat{\mathbf{f}}_{\text{POS}} = \arg \max_{\mathbf{f}} p\left(\mathbf{f}, \widehat{\boldsymbol{\phi}}_{\text{POS}}, \widehat{\boldsymbol{\psi}}_{\text{POS}} | \mathbf{g}\right) \\ \widehat{\boldsymbol{\phi}}_{\text{POS}} = \arg \max_{\boldsymbol{\phi}} p\left(\widehat{\mathbf{f}}_{\text{POS}}, \boldsymbol{\phi}, \widehat{\boldsymbol{\psi}}_{\text{POS}} | \mathbf{g}\right) \\ \widehat{\boldsymbol{\psi}}_{\text{POS}} = \arg \max_{\boldsymbol{\psi}} p\left(\widehat{\mathbf{f}}_{\text{POS}}, \widehat{\boldsymbol{\phi}}_{\text{POS}}, \boldsymbol{\psi} | \mathbf{g}\right) \end{cases}$$

- POS is weaker than JMAP, i.e., JMAP $\Rightarrow$ POS, but POS $\not\Rightarrow$ JMAP.
- How to find a POS? Simply cycle through its defining equations until a stationary point is reached.

Parameter estimation in Bayesian inference problems

Implementing the marginal ML criterion: the EM algorithm.

- Recall the the MML criterion is

$$\begin{aligned}(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}})_{\text{MML}} &= \arg \max_{(\boldsymbol{\phi}, \boldsymbol{\psi})} p(\mathbf{g}|\boldsymbol{\phi}, \boldsymbol{\psi}) \\ &= \arg \max_{(\boldsymbol{\phi}, \boldsymbol{\psi})} \int p(\mathbf{g}, \mathbf{f}|\boldsymbol{\phi}, \boldsymbol{\psi})\, d\mathbf{f}\end{aligned}$$

- Usually, it is infeasible to obtain the marginal likelihood analytically.
- Alternative: use the *expectation-maximization* (EM) algorithm.

Parameter estimation in Bayesian inference problems

- The EM algorithm:

  E-Step: Compute the so-called Q-function. This is the expected value of the logarithm of the complete likelihood function, given the current parameter estimates

$$Q(\boldsymbol{\phi}, \boldsymbol{\psi} | \widehat{\boldsymbol{\phi}}^{(n)}, \widehat{\boldsymbol{\psi}}^{(n)}) = \int p(\mathbf{f}|\mathbf{g}, \widehat{\boldsymbol{\phi}}^{(n)}, \widehat{\boldsymbol{\psi}}^{(n)}) \log p(\mathbf{g}, \mathbf{f}|\boldsymbol{\phi}, \boldsymbol{\psi}) \, d\mathbf{f};$$

  M-Step: Update the parameter estimate according to

$$\left(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}}\right)^{(n+1)} = \arg \max_{(\boldsymbol{\phi}, \boldsymbol{\psi})} Q(\boldsymbol{\phi}, \boldsymbol{\psi} | \widehat{\boldsymbol{\phi}}^{(n)}, \widehat{\boldsymbol{\psi}}^{(n)}).$$

- Under certain (mild) conditions,

$$\lim_{n \to \infty} \left(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}}\right)^{(n)} = \left(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\psi}}\right)_{\text{MML}}$$

Back to the image restoration problem.

- We have a prior $p(\mathbf{f}|\mathbf{h}, \mathbf{v}, \mu)$
- We have an observation model $p(\mathbf{g}|\mathbf{f}, \sigma^2) \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$.
- We have unknown parameters μ, σ^2, $\mathbf{h}$, and $\mathbf{v}$
- Our complete set of unknowns is $(\mathbf{f}, \mu, \sigma^2, \mathbf{h}, \mathbf{v})$
- We need a hyper-prior $p(\mu, \sigma^2, \mathbf{h}, \mathbf{v})$
- It makes sense to assume independence

$$p(\mu, \sigma^2, \mathbf{h}, \mathbf{v}) = p(\mu)\, p(\sigma^2)\, p(\mathbf{h})\, p(\mathbf{v})$$

- We also choose $p(\mu) = k_1$ and $p(\sigma^2) = k_2$, i.e., we will look for ML-type estimates of these parameters.

Reparametrization of the edge variables.

- A natural parametrization of the edge variables uses the locations of those that are equal to 1, which are usually a small minority.
- Let $\boldsymbol{\theta}^h_{(k_h)}$ and $\boldsymbol{\theta}^v_{(k_v)}$ be defined according to

$$\begin{aligned} h_{i,j} &= 1 \Leftrightarrow (i,j) \in \boldsymbol{\theta}^h_{(k_h)} \\ v_{i,j} &= 1 \Leftrightarrow (i,j) \in \boldsymbol{\theta}^v_{(k_v)} \end{aligned}$$

  $\boldsymbol{\theta}^h_{(k_h)}$ contains the locations of the k_h variables $h_{i,j}$ that are set to 1. Similarly for $\boldsymbol{\theta}^v_{(k_v)}$ with respect to the $v_{i,j}$'s.

- Example: if $h_{2,5} = 1$, $h_{6,2} = 1$, $v_{3,4} = 1$, $v_{5,7} = 1$, and $v_{9,12} = 1$, then $k_h = 2$, $k_v = 3$, and

$$\begin{aligned} \boldsymbol{\theta}^h_{(2)} &= [(2,5),(6,2)] \\ \boldsymbol{\theta}^v_{(3)} &= [(3,4),(5,7),(9,12)] \end{aligned}$$

Reparametrization of the edge variables.

- We have two unknown parameter vectors: $\boldsymbol{\theta}^h_{(k_h)}$ and $\boldsymbol{\theta}^v_{(k_v)}$.
- These parameter vectors have unknown dimension, k_h=?, k_v=?
- We have a "model selection problem"
- This justifies another detour: **model selection**.

Returning to our image restoration problem.

- We have two parameter vectors $\boldsymbol{\theta}^h_{(k_h)}$ and $\boldsymbol{\theta}^v_{(k_v)}$ of unknown dimension.
- The natural description length is

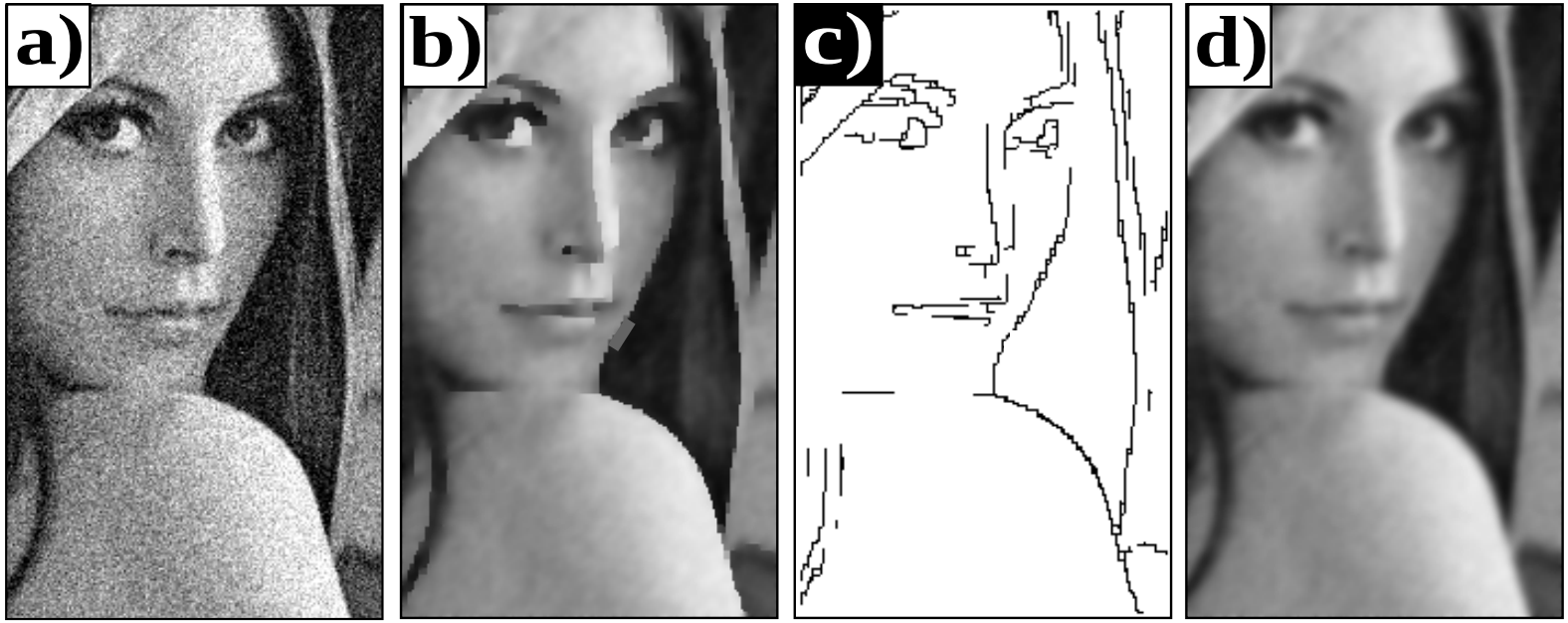
$$\begin{aligned} L\left(\boldsymbol{\theta}^h_{(k_h)}\right) &= k_h(\log M + \log N) \\ L\left(\boldsymbol{\theta}^v_{(k_v)}\right) &= k_v(\log M + \log N) \end{aligned}$$

  where we are assuming the image size is $M \times N$.
- With this MDL "prior" we can now estimate μ, σ^2, $\boldsymbol{\theta}^h_{(k_h)}$, $\boldsymbol{\theta}^v_{(k_v)}$, and, most importantly, $\mathbf{f}$.

Example: discontinuity-preserving restoration

Using the MDL prior for the parameters, and the POS criterion.

(a) Noisy image; (b) discontinuity-preserving restoration; (c) signaled discontinuities; and (d) restoration without preserving discontinuities.



Discontinuity-preserving restoration: implicit discontinuities

- Alternative to explicitly detection/preservation of edges: replace the quadratic potentials by "less aggressive" functions.
- Clique-potentials, for first order auto-model

$$\begin{aligned} V_{\{(i,j),(i,j-1)\}}(f_{i\,j}, f_{i\,j-1}) &= \mu\, \varphi(f_{i\,j} - f_{i\,j-1}) \\ V_{\{(i,j),(i-1,j)\}}(f_{i\,j}, f_{i-1\,j}) &= \mu\, \varphi(f_{i\,j} - f_{i-1\,j}) \end{aligned}$$

  where $\varphi(\cdot)$ is no longer a quadratic function.
- Several $\varphi(\cdot)$'s have been proposed: convex and non-convex.

Discontinuity-preserving restoration: convex potentials

- Generalized Gaussians (Bouman and Sauer [16]), $\varphi(x) = |x|^p$, with $p \in [1, 2]$ (for $p = 2 \Rightarrow$ GMRF).
- Stevenson *et al.* [80] proposed

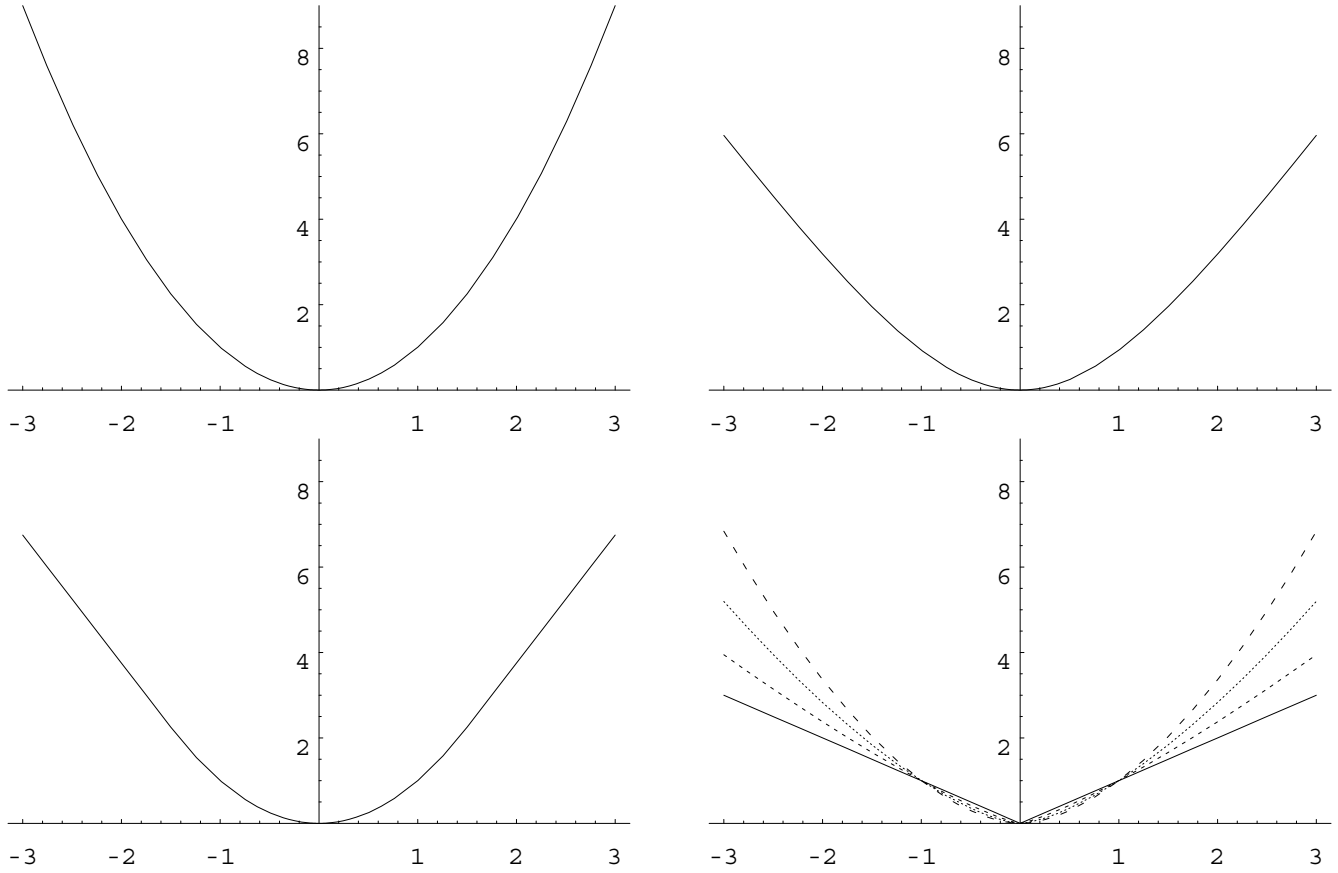
$$\varphi(x) = \begin{cases} x^2 & \Leftarrow |x| < \alpha \\ 2\alpha|x| - \alpha^2 & \Leftarrow |x| < \alpha, \end{cases}$$

- The function proposed by Green [40], $\varphi(x) = 2\alpha^2 \log \cosh(x/\alpha)$.

  Approximately quadratic for small x; linear, for large x.

  Parameter α controls the transition between the two behaviors.

Discontinuity-preserving restoration: convex potentials

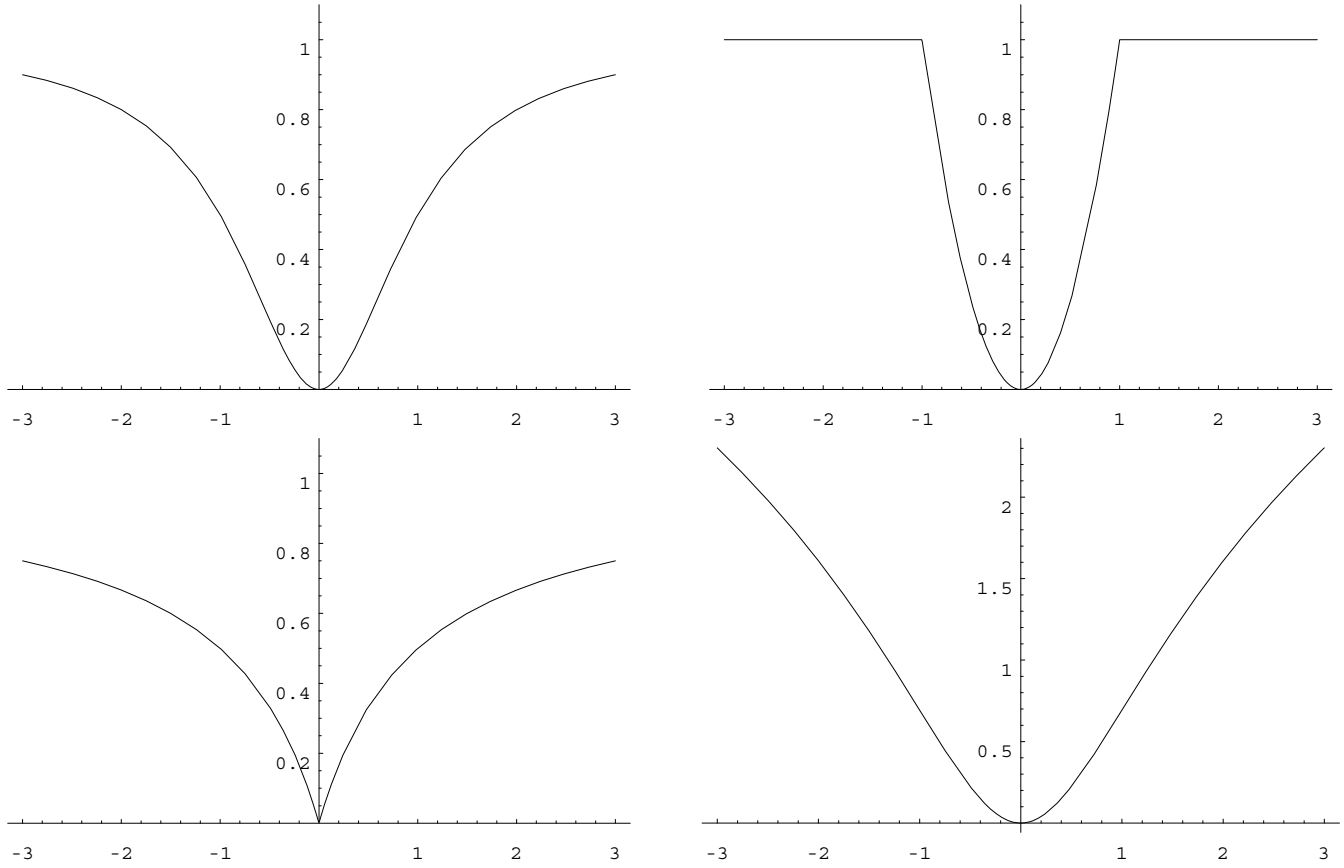


Discontinuity-preserving restoration: non-convex potentials

Radically different from the quadratic: they flatten, for large arguments.

- Blake and Zisserman's $\varphi(x) = (\min\{|x|, \alpha\})^2$ [15], [30]
- The one proposed by Geman and McClure [35]: $\varphi(x) = x^2/(x^2 + \alpha^2)$
- Geman and Reynolds [36] proposed: $\varphi(x) = |x|/(|x| + \alpha)$.
- The one suggested by Hebert and Leahy [45] is $\varphi(x) = \log\left(1 + (x/\alpha)^2\right)$.

Discontinuity-preserving restoration: non-convex potentials



Optimization Problems

- By far the most common criterion in MRF applications is the MAP.
- This requires locating the mode(s) of the posterior

$$\begin{aligned}\widehat{\mathbf{f}}_{\text{MAP}} &= \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{g}) \\ &= \arg\max_{\mathbf{f}} \frac{1}{Z_p(\mathbf{g})} \exp\{-U_p(\mathbf{f}|\mathbf{g})\} \\ &= \arg\min_{\mathbf{f}} U_p(\mathbf{f}|\mathbf{g}),\end{aligned}$$

  where $U_p(\mathbf{f}|\mathbf{g})$ is called the *a posteriori energy*.

- Except in very particular cases (GMRF prior and Gaussian noise) there is no analytical solution.
- Finding a MAP estimate is then a difficult task.

Optimization Problems: Simulated Annealing

- Notice that $U_p(\mathbf{f}|\mathbf{g})$ can be multiplied by any positive constant

$$\begin{aligned}
\arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{g}) &= \arg\max_{\mathbf{f}} \frac{1}{Z_p} \exp\{-U_p(\mathbf{f}|\mathbf{g})\} \\
&= \arg\max_{\mathbf{f}} \frac{1}{Z_p(T)} \exp\left\{-\frac{U_p(\mathbf{f}|\mathbf{g})}{T}\right\} \\
&= \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{g}, T).
\end{aligned}$$

- By analogy with the Boltzman distribution, T is called *temperature*.
- $T \to \infty$, $p(\mathbf{f}|\mathbf{g}, T)$ becomes flat: all configurations equiprobable.
- $T \to 0$, the set of maximizing configurations (denoted Ω_0) gets probability one. Formally

$$\lim_{T \to 0} p(\mathbf{f}|\mathbf{g}, T) = \begin{cases} \frac{1}{|\Omega_0|} & \Leftarrow \quad \mathbf{f} \in \Omega_0 \\ 0 & \Leftarrow \quad \mathbf{f} \notin \Omega_0. \end{cases}$$

Optimization Problems: Simulated Annealing

- *Simulated annealing* (SA): exploits this behavior of $p(\mathbf{f}|\mathbf{g}, T)$
  - Simulate a system whose equilibrium distribution is $p(\mathbf{f}|\mathbf{g}, T)$
  - "Cool" it until the temperature reaches zero.
- Implementation issues of SA:
  - Question: How to simulate a system with equilibrium distribution $p(\mathbf{f}|\mathbf{g}, T)$?

    Answer: Metropolis algorithm or Gibbs sampler.
  - Question: How to "cool it down" without destroying the equilibrium?

    Answer: later.

The Metropolis algorithm.

Simulating a system with equilibrium distribution
$p(\mathbf{f}, T) \propto \exp\{-\frac{U(\mathbf{f})}{T}\}$.

- Starting state $\mathbf{f}(0)$
- Given the current state $\mathbf{f}(t)$, a random "candidate" $\mathbf{c}$ is generated.

  Let $G_{\mathbf{f}(t),\mathbf{c}}$ be the probability of the candidate configuration $\mathbf{c}$, given the current $\mathbf{f}(t)$.
- The candidate $\mathbf{c}$ is accepted with probability $A_{\mathbf{f}(t),\mathbf{c}}(T)$.

  $\mathbf{A}(T) = [A_{\mathbf{f},\mathbf{c}}(T)]$ is the *acceptance matrix.*.
- The new state $\mathbf{f}(t+1)$ only depends on $\mathbf{f}(t)$; this is a Markov chain.

The Metropolis algorithm

- Under certain conditions on $\mathbf{G}$ and $\mathbf{A}(T)$, the equilibrium distribution of is

$$p(\mathbf{f}, T) = \frac{A_{\mathbf{f}_0,\mathbf{f}}(T)}{\sum_{\mathbf{v}\in\Omega} A_{\mathbf{f}_0,\mathbf{v}}(T)}, \quad \text{where } \mathbf{f}_0 \in \Omega_0.$$

- Usual choice

$$A_{\mathbf{f}(t),\mathbf{c}}(T) = \min\left\{1, \exp\left\{\frac{U(\mathbf{f}(t)) - U(\mathbf{c})}{T}\right\}\right\}$$

leading to

$$p(\mathbf{f}, T) \propto \exp\left\{\frac{U(\mathbf{f}_0) - U(\mathbf{f})}{T}\right\} \propto \exp\left\{\frac{-U(\mathbf{f})}{T}\right\}$$

The Gibbs sampler

- Replaces the generation/acceptance mechanism by a simpler one exploiting the Markovianity of $p(\mathbf{f})$
- Current state $\mathbf{f}(t)$
- Choose a site (i.e., an element of $\mathbf{f}(t)$), say $f_i(t)$.
- Generate $\mathbf{f}(t+1)$ by replacing $f_i(t)$ by a random sample of its conditional probability, with respect to $p(\mathbf{f}, T)$. All other elements are unchanged.
- If every site is visited infinitely often, the equilibrium distribution is again $p(\mathbf{f}) \propto \exp\{-\frac{U(\mathbf{f})}{T}\}$

Simulated annealing: cooling

- The temperature evolves according to $T(t)$, called the "cooling schedule".
- The cooling schedule must verify

$$\sum_{t=1}^{\infty} \exp\left\{-\frac{K}{T(t)}\right\} = \infty$$

  where K is a problem-dependent constant

- Best known case:

$$T(t) = \frac{C}{\log(t+1)}$$

  with $C \geq K$.

Iterated conditional modes (ICM) algorithm

- It is a Gibbs sampler at zero temperature.
- The visited site is replaced by the maximizer of its conditional, given the current state of its neighbors.
- Advantage: extremely fast.
- Disadvantage: convergence to local maximum.

  Sometimes, this may not really be a disadvantage.

Implementing the PM and MPM criterion

- Recall: *maximizer of posterior marginals* (MPM)

$$\widehat{\mathbf{f}}_{\mathrm{MPM}} = \left[\arg\max_{f_1} p(f_1|\mathbf{g}) \quad \arg\max_{f_1} p(f_2|\mathbf{g}) \quad \cdots \arg\max_{f_m} p(f_m|\mathbf{g}) \right]^T ;$$

  The *posterior* mean (PM) $\widehat{\mathbf{f}}_{\mathrm{PM}} = E[\mathbf{f}|\mathbf{g}]$.

- Simply simulate (i.e., sample from) $p(\mathbf{f}|\mathbf{g})$ using the Gibbs sampler or the Metropolis algorithm.

- Collect statistics:
  - For the PM, site-wise averages approximate the PM estimate.
  - For the MPM, collect site-wise histograms;
    These histograms are estimates of the marginals $p(f_i|\mathbf{g})$.
    From these (estimated) marginal distributions, the MPM is easily obtained.

The Partition Function Problem

- MRFs are plagued by the difficulty of computing the partition functions.
- This is specially true for parameter estimation.
- Few exceptions: GMRF and Ising fields.
- This issue is dealt with by applying approximation techniques.

Approximating the partition function: Pseudo-likelihood

- Besag's pseudo-likelihood approximation:

$$p(\mathbf{f}) \simeq \prod_{i \in \mathbf{N}} p(f_i | f_{N(i)})$$

- This approximation was used in the example shown above on discontinuity-preserving restoration with CGMRF's and MDL priors

Approximating the partition function: Mean field

- Imported from statistical physics.
- The exact function is approximated by a factored version

$$p(\mathbf{f}) = \frac{\exp\{-U(\mathbf{f})\}}{Z} \simeq \prod_{i \in \mathbf{N}} \frac{\exp\{-U_i^{\mathrm{MF}}(f_i)\}}{Z_i^{\mathrm{MF}}}$$

- The quantity $U_i^{\mathrm{MF}}(f_i)$ is the *mean field local energy*:

$$U_i^{\mathrm{MF}}(f_i) = \sum_{C:\, i \in C} V_C\left(f_i, \{E^{\mathrm{MF}}[f_k] : k \in C\}\right)$$

where

$$E^{\mathrm{MF}}[f_k] = \sum \frac{f_k}{Z_k^{\mathrm{MF}}} \exp\{-U_k^{\mathrm{MF}}(f_k)\}$$

- We replace the neighbors of each site by their (frozen) means.

Mean field approximation (cont.)

- There is a self-referential aspect in the previous equations:
  - To obtain $U_i^{\mathrm{MF}}(f_i)$ we need its neighbors mean values.
  - These, in turn, depend on $E^{\mathrm{MF}}[f_i]$ (since neighborhood relations are symmetrical), thus on $U_i^{\mathrm{MF}}(f_i)$ itself.
- As a consequence, the MF approximation has to be obtained iteratively.
- Alternative: the mean of each site $E^{\mathrm{MF}}[f_i]$ is approximated by the mode: *saddle point approximation*.

Deterministic optimization: Continuation methods

- Continuation methods: the objective function $U(\mathbf{f}|\mathbf{g})$ is embedded in a family

$$\{U(\mathbf{f}|\mathbf{g}, \alpha), \ \ \alpha \in [0, 1]\}$$

  such that

  $U(\mathbf{f}|\mathbf{g}, 0)$ is easily minimizable

  $U(\mathbf{f}|\mathbf{g}, 1) = U(\mathbf{f}|\mathbf{g})$.

- Procedure:
  - Find the minimum of $U(\mathbf{f}|\mathbf{g}, 0)$; this is easy;
  - Track that minimum while α (slowly) increases up to 1.

Deterministic optimization: Continuation methods

- The "tracking" is usually implemented as follows:
  - A discrete set of n values $\{\alpha_0 = 0, \alpha_1, ..., \alpha_t, ..., \alpha_{n-1}, \alpha_n = 1\} \subset [0, 1]$ is chosen;
  - for each α_t, $U(\mathbf{f}|\mathbf{g}, \alpha_t)$ is minimized by some local iterative technique.
  - This iterative process is initialized at the previously obtained minimum for α_{t-1}.
- Writing $T = -\log \alpha$ reveals that simulated annealing shares some of the spirit of continuation algorithms.

  Simulated annealing can be called a "stochastic continuation method".

Continuation methods: Mean field annealing (MFA)

- MFA is a deterministic surrogate of (stochastic) simulated annealing.
- $p(\mathbf{f}|\mathbf{g}, T)$ is replaced by its MF approximation.
- Computing the MF approximation $\Leftrightarrow$ finding the MF values.

  The fact that these must obtained iteratively is exploited to insert its computation into a continuation method

Continuation methods: Mean field annealing (MFA)

- For $T \to \infty$, $p(\mathbf{f}|\mathbf{g}, T)$ and its MF approximation are uniform.

  The mean field is trivially obtainable.

- At (finite) temperature T_t, the mean field values $E_t^{\text{MF}}[f_k|\mathbf{g}, T_t]$ are obtained iteratively.

- This iterative process is initialized at the previous mean field values $E_{t-1}^{\text{MF}}[f_k|\mathbf{g}, T_{t-1}]$

- As $T(t) \to 0$, the MF approximation converges to a distribution concentrated on its global maxima.

- Alternatively, temperature descent is stopped at $T = 1$.

  This yields a MF approximation of $p(\mathbf{f}|\mathbf{g}, T = 1) = p(\mathbf{f}|\mathbf{g})$

  Mean field values are (approximate) PM estimates.

Continuation methods: Simulated tearing (ST)

- Uses the following family of functions

$$\{U(\mathbf{f}|\mathbf{g},\alpha) = U(\mathbf{f}|\alpha\,\mathbf{g}),\ \ \alpha \in [0,1]\}\,.$$

- Obviously, $U(\mathbf{f}|\mathbf{g},1) = U(\mathbf{f}|\mathbf{g})$
- This method is adequate when $U(\mathbf{f}|0)$ is easily minimizable.
- This is the case of most discontinuity-preserving MRF priors, because for $\mathbf{g} \simeq 0$, the potentials have convex behavior.
- The example shown above, of discontinuity-preserving restoration with CGMRF and MDL priors, uses this continuation method.

Important topics not covered:

- Specific discrete-state MRF's: Ising, auto-logistic, auto-binomial...
- Multi-scale MRF models.
- Causal MRF models.
- Closer look at applications (see references).

Some references (this is not an exhaustive list):

Fundamental Bayesian theory. See accompanying text and the many references therein.

Compound inference. General concepts: [7], [74]. In computer vision/image analysis/pattern recognition: [4], [44], [65]. The multivariate Gaussian case, from a signal processing perspective: [76].

Random fields on graphs: [41], [42], [71], [79] (and references therein), and [81].

Markov random fields on regular graphs:

Seminal papers: [33], [11].

Earlier work: [82], [75], [85], [86], [9], [52].

Books (these are good sources for further references): [19], [62], [83], [42]. See also [41].

Influential papers on MRFs for image analysis and computer vision: [11], [18], [20], [21], [23], [24], [25], [26], [33], [49], [61], [65], [68], [85], [86].

Compound Gauss-Markov Random fields and applications: [28], [48], [49]. [90].

Parameter estimation: [6], [55], [28], [39], [50], [56], [64], [67], [84], [91], [94], [5], and further references in [62].

Specific references on the EM algorithm and its applications:

Fundamental work: [22], [63], [87]

Some applications: [45], [54], [58], [59], [91].

Model Selection (including MDL and its applications): [8], [28], [29], [32], [57], [60], [51] (and references therein), [73], [77], [93].

Discontinuity-preserving priors: [15], [16], [30], [35], [36], [40], [45], [80].

Pseudo-likelihood approximation: [34], [38], [41], [10].

Mean field approximation:

Statistical physics: [17], [70].

In MRF's literature: [14], [88], [30], [31], [90], [91], [92], [94].

Simulated annealing (including the Gibbs sampler and the Metropolis algorithm): [1], [3], [2], [12], [33], [37], [43], [47], [53], [66], [69], [83].

Iterated conditional modes (ICM): [11].

Mean field annealing: [13], [14], [30], [31], [46], [78], [89], [90].

Other continuation methods (including "simulated tearing": [15], [27], [28], [72].

References

[1] E. Aarts and P. van Laarhoven. Statistical cooling: A general approach to combinatorial optimization problems. *Philips Journal of Research*, 40(4):193–226, 1985.

[2] E. Aarts and P. van Laarhoven. *Simulated annealing : theory and applications*. Kluwer Academic Publishers, Dordrecht (Netherlands), 1987.

[3] E. Aarts and P. van Laarhoven. Simulated annealing: A pedestrian review of the theory and some applications. In P. Devijver and J. Kittler, editors, *Pattern Recognition Theory and Applications – NATO Advanced Study Institute*, pages 179–192. Springer Verlag, 1987.

[4] K. Abend, T. Harley, and L. Kanal. Classification of binary random patterns. *IEEE Transactions on Information Theory*, 11, 1965.

[5] G. Archer and D. Titterington. On some Bayesian/Regularization methods for image restoration. *IEEE Transactions on Image Processing*, IP-4(7):989–995, July 1995.

[6] N. Balram and J. Moura. Noncausal Gauss-Markov random fields: Parameter structure and estimation. *IEEE Transactions on Information Theory*, IT-39(4):1333–1355, July 1993.

[7] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1980.

[8] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, Chichester (UK), 1994.

[9] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36(2):192–225, 1974.

[10] J. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3):616–618, 1977.

[11] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48(3):259–302, 1986.

[12] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10:3–66, 1995.

[13] G. Bilbro and W. Snyder. Range image restoration using mean field annealing. In *Advances in Neural Network Information Processing Systems*, San Mateo, CA, 1989. Morgan-Kaufman.

[14] G. Bilbro, W. Snyder, S. Garnier, and J. Gault. Mean field annealing: A formalism for constructing GNC-Like algorithms. *IEEE Transactions on Neural Networks*, 3(1):131–138, January 1992.

[15] A. Blake and A. Zisserman. *Visual Reconstruction.* M.I.T. Press, Cambridge, M.A., 1987.

[16] C. Bouman and K. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Transactions on Image Processing*, IP-2:296–310, January 1993.

[17] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, Oxford, 1987.

[18] R. Chellappa. Two-dimensional discrete Gaussian Markov random field models for image processing. In L. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition.* Elsevier Publ., 1985.

[19] R. Chellappa and A. Jain (Editors). *Markov Random Fields: Theory and Applications.* Academic Press, San Diego, CA, 1993.

[20] P. Chou and C. Brown. The theory and practice of Bayesian image labeling. *International Journal of Computer Vision*, 4:185–210, 1990.

[21] F. Cohen and D. Cooper. Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9:195–219, 1988.

[22] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[23] H. Derin. The use of Gibbs distributions in image processing. In I. Blake and H. Poor, editors, *Communications and Networks: A Survey of Recent Advances*, pages 266–298, New-York, 1986. Springer-Verlag.

[24] H. Derin and H. Elliot. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, 1987.

[25] H. Derin and P. Kelly. Discrete-index Markov-type random processes. *Proceedings of the IEEE*, 77(10):1485–1510, October 1989.

[26] R. Dubes and A. K. Jain. Random field models for image analysis. *Journal of Applied Statistics*, 6:131–164, 1989.

[27] M. Figueiredo and J. Leitão. Simulated tearing: an algorithm for discontinuity preserving visual surface reconstruction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR'93*, pages 28–33, New York, June 1993.

[28] M. Figueiredo and J. Leitão. Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle. *IEEE Transactions on Image Processing*, IP-6(8):1089–1102, August 1997.

[29] M. Figueiredo, J. Leitão, and A. K. Jain. Adaptive B-splines and boundary estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR'97*, pages 724–729, San Juan (PR), 1997.

[30] D. Geiger and F. Girosi. Parallel and deterministic algorithms from MRF's: Surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(5):401–412, May 1991.

[31] D. Geiger and A. Yuille. A common framework for image segmentation. *International Journal of Computer Vision*, 6(3):227–243, 1991.

[32] A. Gelfand and D. Dey. Bayes model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society B*, 56:501–514, 1994.

[33] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[34] S. Geman and C. Graffigne. Markov random field image models and their applications to computer vision. *Proceedings of the International Congress of Mathematicians*, pp. 1496–1517. 1987.

[35] S. Geman, D. McClure, and D. Geman. A nonlinear filter for film restoration and other problems in image processing. *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, 54(4):281–289, July 1992.

[36] S. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-14(3):367–383, March 1992.

[37] B. Gidas. The Langevin equation as a global minimization algorithm. In E. Bienenstock, F. Fogelman Soulié, and G. Weisbuch, editors, *Disordered Systems and Biological Organization – NATO Advanced Study Institute*, pages 321–326. Springer Verlag, 1986.

[38] B. Gidas. Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. In W. Fleming and P. Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory, and Applications*, pages 129–145. Springer Verlag, New York, 1988.

[39] B. Gidas. Parameter estimation for Gibbs distributions from partially observed data. *Annals of Statistics*, 2(1):142–170, 1992.

[40] P. Green. Bayesian reconstruction from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*, MI-9(1):84–93, March 1990.

[41] U. Grenander. *General Pattern Theory: A Mathematical Study of Regular Structures*. Oxford University Press, Oxford, 1993.

[42] X. Guyon. *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer Verlag, N. York, 1995.

[43] B. Hajek. A tutorial survey of theory and applications of simulated annealing. In *Proceedings of the* 24^{th} *Conference on Decision and Control*, pages 755–760, Fort Lauderdale (FL), 1985.

[44] R. Haralick. Decision making in context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5:417–428, 1983.

[45] T. Hebert and R. Leahy. A generalized EM algorithm for 3D Bayesian reconstruction from poisson data using Gibbs priors. *IEEE Transactions on Medical Imaging*, MI-8:194–202, 1989.

[46] H. Hiriyannaiah, G. Bilbro, W. Snyder, and R. Mann. Restoration of piecewise constant images by mean field annealing. *Journal of the Optical Society of America*, 6(12):1901–1911, December 1989.

[47] M. Hurn and C. Jennison. Multiple-site updates in maximum a posteriori and marginal posterior modes image estimation. In K. Mardia and G. Kanji, editors, *Advances in Applied Statistics: Statistics and Images 1*, pages 155–186. Carfax Publishing, 1993.

[48] F. Jeng and J. Woods. Image estimation by stochastic relaxation in the compound Gaussian case. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing – ICASSP'88*, pages 1016–1019, New York, 1988.

[49] F. Jeng and J. Woods. Compound Gauss-Markov random fields for image estimation. *IEEE Transactions on Signal Processing*, SP-39:683–697, March 1991.

[50] V. Johnson, W. Wong, X. Hu, and C. Chen. Aspects of image restoration using gibbs priors: Boundary modelling, treatment of blurring, and selection of hyperparameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:412–425, 1990.

[51] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:733–795, 1995.

[52] R. Kinderman and J. Snell. *Markov Random Fields and their Applications*. American Mathematical Society, Providence (R.I.), 1980.

[53] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[54] R. Lagendijk, J. Biemond, and D. Boekee. Identification and restoration of noisy blurred images using the expectation-maximization algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(7):1180–1191, July 1990.

[55] S. Lakshmanan and H. Derin. Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(8):799–813, August 1989.

[56] S. Lakshmanan and H. Derin. Valid parameter space for 2D Gaussian Markov random fields. *IEEE Transactions on Information Theory*, 39(2):703–709, March 1993.

[57] D. Langan, J. Modestino, and J. Zhang. Cluster validation for unsupervised stochastic model-based image segmentation. *IEEE Transactions on Image Processing*, 7:180–195, 1998.

[58] K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Transactions on Medical Imaging*, 9:439–446, 1991.

[59] K. Lay and A. Katsaggelos. Blur identification and image restoration based on the EM algorithm. *Optical Engineering*, 29(5):436–445, May 1990.

[60] Y. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3:73–102, 1989.

[61] S. Li. Invariant surface segmentation through energy minimization with discontinuities. *International Journal of Computer Vision*, 5(2):161–194, 1990.

[62] S. Z. Li. *Markov random field modelling in computer vision*. Springer Verlag, Tokyo, 1995.

[63] R. Little and D. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

[64] D. MacKay. Hyperparameters: Optimize, or integrate out? In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 43–60, Dordrecht, 1996. Kluwer.

[65] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[66] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[67] A. Mohammad-Djafari. Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems. In *Proceedings of the IEEE International Conference on Image Processing – ICIP'96*, volume II, pages 473–476, Lausanne, 1996.

[68] J. Moura and N. Balram. Recursive structure of noncausal Gauss-Markov random fields. *IEEE Transactions on Information Theory*, IT-38(2):334–354, March 1992.

[69] R. Otten and L. Ginneken. *The Annealing Algorithm*. Kluwer Academic Publishers, Boston, 1989.

[70] G. Parisi. *Statistical Field Theory*. Addison Wesley Publishing Company, Reading, Massachusetts, 1988.

[71] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, CA, 1988.

[72] A. Rangarajan and R. Chellappa. Generalized graduated non-convexity algorithm for maximum a posteriori image estimation. In *Proceedings of the* 9^{th} *IAPR International Conference on Pattern Recognition – ICPR'90*, pages 127–133, Atlantic City, 1990.

[73] J. Rissanen. *Stochastic Complexity in Stastistical Inquiry*. World Scientific, Singapore, 1989.

[74] C. Robert. *The Bayesian Choice: A Decision Theoretic Motivation*. Springer-Verlag, New York, 1994.

[75] Y. Rosanov. On Gaussian fields with given conditional distributions. *Theory of Probability and Its Applications*, XII:381–391, 1967.

[76] L. Scharf. *Statistical Signal Processing*. Addison Wesley, Reading, Massachusetts, 1991.

[77] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[78] P. Simic. Statistical mechanics as the underlying theory of "elastic" and "neural" optimisations. *Network*, 1:89–103, 1990.

[79] P. Smythe. Belief networks, hidden Markov models, and Markov random fields: A unifying view. *Pattern Recognition Letters*, 18:1261–1268, 1997.

[80] R. Stevenson, B. Schmitz, and E. Delp. Discontinuity-preserving regularization of inverse visual problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-24(3):455–469, March 1994.

[81] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester, UK, 1990.

[82] P. Whittle. On the stationary process in the plane. *Biometrika*, 41:434–449, 1954.

[83] G. Winkler. *Image analysis, random fields, and dynamic Monte Carlo systems*. Springer-Verlag, Berlin, 1995.

[84] C. Won and H. Derin. Unsupervised segmentation of noisy and textured images using Markov random fields. *Computer Vision, Graphics, and Image Processing (CVGIP): Graphical Models and Image Processing*, 54(4):308–328, 1992.

[85] J. Wong. Two-dimensional random fields and the representation of images. *SIAM Journal of Applied Mathematics*, 16(4), 1968.

[86] J. Woods. Two-dimensional discrete Markovian fields. *IEEE Transactions on Information Theory*, IT-18(2):232–240, March 1972.

[87] C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, vol. 11, 95–103, 1983.

[88] C. Wu and P. Doerschuk. Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):275–293, March 1995.

[89] A. Yuille. Generalized deformable models, statistical physics, and the matching problem. *Neural Computation*, 2:1–24, 1990.

[90] J. Zerubia and R. Chellappa. Mean field annealing using compound Gauss-Markov random fields for edge detection and image estimation. *IEEE Transactions on Neural Networks*, 4(4):703–709, July 1993.

[91] J. Zhang. The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE Transactions on Image Processing*, IP-2(1):27–40, January 1993.

[92] J. Zhang. The convergence of mean field procedures for MRF's. *IEEE Transactions on Image Processing*, IP-5(12):1662–1665, December 1996.

[93] J. Zheng and S. Bolstein. Motion-based object segmentation and estimation using the MDL principle. *IEEE Transactions on Image Processing*, IP-2(9):1223–1235, September 1995.

[94] Z. Zhou, R. Leahy, and J. Qi. Approximate maximum likelihood hyperparameter estimation for Gibbs priors. *IEEE Transactions on Image Processing*, 6(6):844–861, June 1997.