Lecture 32

Prior models

Goal: Write a density \( \pi_{pr}(x) \) with the following property: If \( E \) is a coll. of expectable vec's \( x \) representing possible realizations of \( X \), and \( U \) is a collection of unexpectable ones, we should have

\[
\pi_{pr}(x) \geq \pi_{pr}(x') \quad \text{when } x \in E, x' \in U
\]

ie, the prior prob. distr. should be concentrated on those values of \( x \) that we expect to see and assign a clearly higher prob. to them than those that we do not expect to see.

We will focus on Gaussian priors, which are the most commonly used in statistical IP's.

Why use Gaussian priors?

1. They are easy to handle
2. Due to the central limit thm., Gaussian densities are often very good approx's to inherently non-Gaussian distr's when the observation is physically based on a large number of mutually indep. random events.

Def 1: Let \( x_0 \in \mathbb{R}^n \) and \( \Gamma \in \mathbb{R}^{n \times n} \) be a sym. pos.

Def mix. A Gaussian \( n \)-variate random variable \( X \) with mean \( x_0 \) and covariance \( \Gamma \) is a random var. with prob. density

\[
\pi(x) = \frac{1}{(2\pi)^{n/2} |\Gamma|^{1/2}} \exp \left( -\frac{1}{2} (x-x_0)^T \Gamma^{-1} (x-x_0) \right)
\]

where \( |\Gamma| = \text{det}(\Gamma) \). We say the notation \( X \sim \mathcal{N}(x_0, \Gamma) \).
Def 2. A random variable $X$ is Gaussian if
\[ E \{ \exp(-i \mathbf{z}^T \mathbf{X}) \} = \exp(-i \mathbf{z}^T \mathbf{X}_0 - \frac{1}{2} \mathbf{z}^T \mathbf{\Gamma} \mathbf{z}) \]
where $\mathbf{X}_0 \in \mathbb{R}^n$ and $\mathbf{\Gamma} \in \mathbb{R}^{n \times n}$ is pos. semi-def., i.e., $\mathbf{\Gamma} \succeq \mathbf{0}$.

There are closed formulas for the conditional means & covariances of Gaussian random variables.

Ex: Assume we have a linear model with additive noise, $Y = \mathbf{A} \mathbf{X} + \mathbf{E}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known $m \times n$, and $\mathbf{X} : \Omega \rightarrow \mathbb{R}^n$, $\mathbf{Y}, \mathbf{E} : \Omega \rightarrow \mathbb{R}^m$ are random variables. Assume $\mathbf{X}$ and $\mathbf{E}$ are mutually independent Gaussian random variables with probability densities
\[ \Pi_{\mathbf{pr}}(\mathbf{x}) \propto \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{\Gamma}^{-1}_{\mathbf{pr}} (\mathbf{x} - \mathbf{x}_0)\right) \]
\[ \Pi_{\mathbf{pr}}(\mathbf{e}) \propto \exp\left(-\frac{1}{2} (\mathbf{e} - \mathbf{e}_0)^T \mathbf{\Gamma}^{-1}_{\text{noise}} (\mathbf{e} - \mathbf{e}_0)\right) \]

Then by Bayes' formula
\[ \Pi(\mathbf{x} | y) = \frac{\Pi_{\mathbf{pr}}(\mathbf{x}) \Pi_{\text{noise}}(y - \mathbf{A} \mathbf{x})}{\int \Pi_{\mathbf{pr}}(\mathbf{x}) \Pi_{\text{noise}}(y - \mathbf{A} \mathbf{x}) \, d\mathbf{x}} \]
\[ \propto \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \mathbf{\Gamma}^{-1}_{\mathbf{pr}} (\mathbf{x} - \mathbf{x}_0) - \frac{1}{2} (y - \mathbf{A} \mathbf{x} - \mathbf{e}_0)^T \mathbf{\Gamma}^{-1}_{\text{noise}} (y - \mathbf{A} \mathbf{x} - \mathbf{e}_0)\right) \]

Suppose now we take the Gaussian white noise prior where the covariance $\mathbf{\Gamma}$ is proportional to the identity and the mean of the prior is $\mathbf{0}$, i.e., $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Also assume the noise is white noise, i.e., $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. 


Then one can calculate the posterior mean explicitly as (using thm 3.7, Kaipio & Somersalo):

$$\hat{x} = \hat{x} = \mathbf{A}^T(\mathbf{y}^2 \mathbf{AA}^T + \sigma^2 I)^{-1} \mathbf{y}$$

or, letting \(\alpha = \sigma^2 / \mathbf{y}^2\)

$$\hat{x} = \mathbf{A}^T (\mathbf{AA}^T + \alpha I)^{-1} \mathbf{y}$$

by SVD = \((\mathbf{AA}^T + \alpha I)^{-1} \mathbf{A}^T \mathbf{y}\)

Recall, this is the T-reg’z’d soln of $\mathbf{Ax} = \mathbf{y}$ with reg. pav. \(\alpha\).

So \(\alpha\) is the ratio of the noise variance to prior variance!

We can calculate the posterior density from Bayes’ thm for IPR’s:

$$\Pi_x(x | y) = \Pi_{\text{pr}}(x) \Pi(y | x)$$

$$= \exp\left( -\frac{1}{2 \mathbf{y}^2} x^T \mathbf{A} \mathbf{y} - \frac{1}{2 \sigma^2} (\mathbf{y} - \mathbf{A} x)^T (\mathbf{y} - \mathbf{A} x) \right)$$

$$= \exp\left\{ -\frac{1}{2 \sigma^2} (\| \mathbf{Ax} - \mathbf{y} \|^2 + \frac{\sigma^2}{\mathbf{y}^2} \| x \|^2) \right\}$$

The MAP estimate is given by

arg max \(\Pi_x(x | y)\)

\(\Pi_x(x | y)\) has a max when the exponent has a min.

Hence \(x_{MAP}\) is equiv. to the T-reg’z’d least-squares soln.

Furthermore, the likelihood func. in this case is given by (see lecture 31)

$$\Pi(y | x) = \Pi_{\text{noise}}(\mathbf{y} - \mathbf{A} x) = \exp\left( -\frac{1}{2 \sigma^2} \| \mathbf{y} - \mathbf{A} x \|^2 \right).$$
Maximizing the likelihood func. \( P(y|x) = \exp\left(-\frac{1}{2\alpha^2} \|y-Ax\|^2\right) \) is how we find the estimate \( \hat{x}_{ML} \), but \( P(y|x) \) has a maximum when \( \|y-Ax\|^2 \) is minimized, so this is equivalent to unregularized least squares!

Interpreting the Posterior Distribution

We have looked at single estimates (\( \hat{x}_{MAP}, \hat{x}_{EM} \) and \( \hat{x}_{ML} \)), but these have some serious shortcomings. (See example 110, sec. 3.6 of Kaipio & Somersalo.)

Alternative idea: Instead of evaluating the prob. density at given pts, let the density itself determine a set of pts, a sample, that well supports the distribution. These sample pts can then be used for approx. integration.

One ex. of a random sampling mtd is the class of Markov chain Monte Carlo techniques (MCMC).

General idea: MCMC mtds are systematic ways of generating a set of samples such that the following holds: Let \( f \) be a scalar or vector-val'd measurable func integrable over \( \mathbb{R}^n \) wrt the measure \( \mu \), i.e., \( f \in L^1(\mu(dx)) \). We wish to find an ensemble of samples \( \{x_1, \ldots, x_N\} \) distributed according to the prob. distr \( \mu \) s.t.

\[ \int_{\mathbb{R}^n} f(x) \mu(dx) = \mathbb{E}[f(X)] \approx \frac{1}{N} \sum_{j=1}^{N} f(x_j). \]
A discrete time stochastic process is an ordered set \( \{X_j\}_{j=1}^{\infty} \) of random variables \( X_j \in \mathbb{R}^n \).

Let \( \mathcal{B} = \mathcal{B}(\mathbb{R}^n) \) denote the Borel sets over \( \mathbb{R}^n \). A mapping \( P : [\mathbb{R}^n \times \mathcal{B}] \rightarrow [0,1] \) is called a probability transition kernel if:

1. for each \( B \in \mathcal{B} \), the mapping \( \mathbb{R}^n \rightarrow [0,1] \), \( x \mapsto P(x,B) \) is a measurable function.
2. for each \( x \in \mathbb{R}^n \), the mapping \( \mathcal{B} \rightarrow [0,1] \), \( B \mapsto P(x,B) \) is a probability distribution.

A time-homogeneous Markov chain with the transition kernel \( P \) is a stochastic process \( \{X_j\}_{j=1}^{\infty} \) with the properties:

\[
\begin{align*}
\mu |_{X_{j+1}} (B_{j+1} | x_1, \ldots, x_j) &= \mu |_{X_{j+1}} (B_{j+1} | x_j) \\
&= P(x_j, \tilde{B}_{j+1}) (1)
\end{align*}
\]

and

\[
\begin{align*}
\mu |_{X_{j+1}} (B_{j+1} | x_1, \ldots, x_j) &= \mu |_{X_{j+1}} (B_{j+1} | x_j) \\
&= P(x_j, \tilde{B}_{j+1}) (2)
\end{align*}
\]

\( (1) \Rightarrow \) the probability for \( X_{j+1} \in B_{j+1} \) conditioned on observations \( X_1 = x_1, \ldots, X_j = x_j \) equals the probability conditioned on \( X_j = x_j \) alone.

\( (2) \Rightarrow \) time is homogeneous in the sense that the dependence of adjacent moments does not vary in time. Note that the kernel \( P \) does not depend on time \( j \).
More generally, we define the transition kernel that propagates \( k \) steps forward in time, setting

\[
P^{(k)}(x_j, B_{j+k}) = \mu_{x_j}(B_{j+k} \mid x_j)
\]

\[
= \int_{\mathbb{R}^n} P(x_{j+k-1}, B_{j+k}) P^{(k-1)}(x_j, dx_{j+k-1}),
\]

for \( k \geq 2 \), and

\[
P^{(1)}(x_j, B_{j+1}) = P(x_j, B_{j+1}).
\]

The two most common techniques for constructing trans. kernels with the correct properties are

1. The Metropolis-Hastings algorithm
2. The Gibbs sampler

These are described in detail in \( K+S \).