§ 4 Introduction to Bayesian Inference

fundamental problem of statistical inference

given a set of observed realizations of a random variable X, $\{x_1, x_2, \ldots, x_n\}$, $x_i \in X$, we want to infer the underlying probability distribution that gives rise to $\{x_1, x_2, \ldots, x_n\}$

- parametric approach: it is assumed that underlying probability distribution giving rise to data can be described in terms of a number of parameters; inference problem: infer parameters / their distribution
- non parametric approach: no assumptions are made on form of underlying probability distribution;

the Bayesian framework

ingredients:
- distribution for data conditioned on unknown parameters
- probability distribution to model information about all unknown parameters

likelihood $\pi_{LIKE}(y|x)$: density that can be viewed as a function of unknown parameters x for a fixed $Y = y$ it is called sampling density of Y when x is fixed

prior $\pi_{PRIOR}(x)$: we assume that x is modelled by a random variable X with density $\pi(x)$, then $\pi(x)$ is called the prior density

posterior $\pi_{POST}(x|y)$: probability density of X for a given $Y = y$; solution of Bayesian inverse problem

Bayes' theorem

posterior $\propto$ likelihood $\times$ prior

construction of likelihood and prior makes use of marginalization and conditioning

conditioning can be used to take into consideration one unknown at a time, pretending the others are fixed

for example, if we want a joint density of x and y, we can write the density of y for a fixed x and then deal with the density of x alone

$$\pi(x,y) = \pi(y|x)\,\pi(x)$$

marginalization can be used to eliminate variables from model that are of no interest by integrating them out

for example, if we have a joint density $\pi(x,y,z)$ but we are not interested in z, we can marginalize it as follows

$$\pi(x,y) = \int_U \pi(x,y,z)\,dz$$

the parameter that is not of interest is referred to as "noise" or a "nuisance" parameter

construction of likelihood

we view the likelihood density as a probability density from which, presumably, the data is generated; in the Bayesian setting, the associated parameters are viewed as realization of random variables;

likelihood can be thought of as answering the following question: if we knew all model parameters x and all parameters defining the data, how would the measurements be distributed?

most common sources of deviations of data from predictions of observational model

(i) measurement noise in the data
- the probability density of the noise can depend on unknown parameters

(ii) incompleteness of the observation model
- includes errors due to discretization, model reduction, and more generally, all the shortcomings of a computational model

assume that $x \in \mathbb{R}^n$ is the unknown of primary interest and that the observable quantity $y \in \mathbb{R}^m$

is ideally related to x through a functional dependence: the ideal deterministic model is given by $y = g(x)$, $g: \mathbb{R}^n \to \mathbb{R}^m$

the available measurements, however, are corrupted by noise attributed to, e.g., external sources or to instabilities in measuring device; they do not depend on x

assuming the data y is perturbed by random additive noise H, we arrive at statistical model

$$Y_{OBS} = g(X) + H = Y + H$$

in the likelihood modeling we pretend to have realizations of X and the task is to construct the distribution of Y_{OBS}

distribution of error: $H \sim \pi_{NOISE}(\eta)$

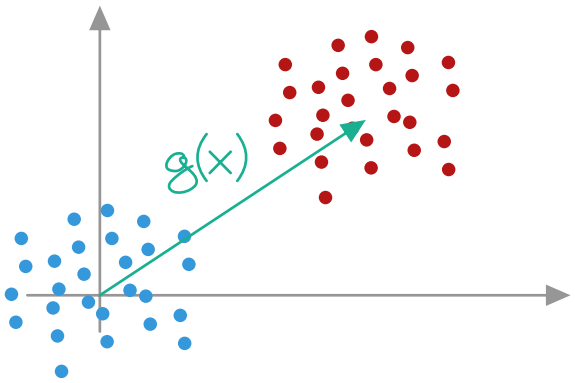
since we assume that noise H does not depend on X, fixing $X = x$ does not change probability distribution of H

$$\pi(\eta \mid x) = \pi_{NOISE}(\eta)$$

on the other hand, if $X = x$ is fixed, the only source of randomness in Y_{OBS} is H

$$\pi_{LIKE}(y_{OBS} \mid x) = \pi_{NOISE}(y_{OBS} - g(x))$$

randomness in Y_{OBS} is randomness of noise H translated by $g(x)$



$$\pi_{\text{NOISE}}(\eta) = \pi_{\text{NOISE}}(\eta \mid \theta)$$

we obtain $\pi_{\text{LIKE}}(y_{\text{OBS}} \mid x, \theta) = \pi_{\text{NOISE}}(y_{\text{OBS}} - g(x) \mid \theta)$

example: let H be zero mean Gaussian with unknown variance τ^2, $H \sim \mathcal{N}(0, \tau^2 I_n)$, $I_n = \text{diag}(1, \dots, 1) \in \mathbb{R}^{n,n}$

the corresponding likelihood model is

$$\pi_{\text{LIKE}}(y_{\text{OBS}} \mid x, \tau^2) = ((2\pi)^{n/2} \tau^n)^{-1} \exp(-(2\tau^2)^{-1} \| y_{\text{OBS}} - g(x) \|^2)$$

with $\theta = \tau^2$

if we assume we know τ^2, we write

$$\pi_{\text{LIKE}}(y_{\text{OBS}} \mid x) \propto \exp(-(2\tau^2)^{-1} \| y_{\text{OBS}} - g(x) \|^2)$$

example: $y_{\text{OBS}}(s) = \int_{-1}^{2} \ker_\tau(s-t) x(t)\, dt + \eta(s), \quad 0 \le s \le 1$

$$\ker_\tau(s-t) = (2\pi\tau^2)^{-1/2} \exp(-(2\tau^2)^{-1}(s-t)^2)$$

midpoint rule for discretization; subdivide $[0,1]$ into intervals $[ih, (i+1)h] \subset \mathbb{R}$, $i = 1, \dots, n$, $h = 1/n$

$$Y_{\text{OBS}} = (y_{\text{OBS}}(s_1), \dots, y_{\text{OBS}}(s_i), \dots, y_{\text{OBS}}(s_n))$$

$$X = (x(t_1), \dots, x(t_i), \dots, x(t_n))$$

discrete deconvolution problem

$$Y_{\text{OBS}} = KX + H$$

with $K = [k_{ij}]_{i,j=1}^{n,n}$, $k_{ij} = h\, \ker_\gamma((i-j)h)$

we assume

$$H \sim \mathcal{N}(0, \gamma^2 I_n),\ I_n = \mathrm{diag}(1,\dots,1) \in \mathbb{R}^{n,n}$$

with this the likelihood is given by

$$\begin{aligned}\pi_{\text{LIKE}}(y_{\text{OBS}} \mid x) &= \pi_{\text{NOISE}}(y_{\text{OBS}} - Kx) \\ &= \mathcal{N}(y_{\text{OBS}} - Kx, \gamma^2 I_n)\end{aligned}$$

the Bayesian solution of inverse problem is by virtue of Bayes formula given by

$$\begin{aligned}\pi_{\text{POST}}(x \mid y_{\text{OBS}}) &\propto \pi_{\text{LIKE}}(y_{\text{OBS}} \mid x)\, \pi_{\text{PRIOR}}(x) \\ &= \mathcal{N}(y_{\text{OBS}} - Kx, \gamma^2 I_n)\, \pi_{\text{PRIOR}}(x)\end{aligned}$$

if we assume multiplicative noise, the statistical model is given by

$$Y_{\text{OBS}} = g(X) \odot H = Y \odot H$$

the likelihood becomes

$$\pi_{\text{LIKE}}(y_{\text{OBS}} \mid x) = \pi_{\text{NOISE}}(y_{\text{OBS}} \oslash g(x)) \Big/ \prod_{i=1}^{m} g_i(x)$$

<u>construction of prior</u>

the prior density expresses what we know / are certain of / believe about the unknown variable of interest prior to taking measurements into account

general goal in designing priors:

if $\in$ is a expectable vectors x, and $\mathcal{U}$ is a collection of unexpectable vectors $\tilde{x}$, then

$$\pi_{\text{PRIOR}}(x) \gg \pi_{\text{PRIOR}}(\tilde{x}) \quad \text{when} \quad x \in \in,\ \tilde{x} \in \mathcal{U}$$

impulse prior densities

assumption: to be reconstructed $x \in \mathbb{R}^n$ is sparse / has few nonzero entries / is localized

option (i): ℓ^1-prior

$$\pi_{\text{PRIOR}}(x) := \left(\frac{\alpha}{2}\right)^n \exp(-\alpha \|x\|_1)$$

with ℓ^1-norm $\|x\|_1 := \sum_{i=1}^n |x_i|$

other option

$$\pi_{\text{PRIOR}}(x) \propto \exp\left(-\alpha \sum_{i=1}^n |x_i|^p\right), \quad 0 < p < 1$$

option (ii): Cauchy density

$$\pi_{\text{PRIOR}}(x) := \left(\frac{\alpha}{\pi}\right)^n \prod_{i=1}^n (1 + \alpha^2 x_i^2)^{-1}$$

option (iii): entropy density; let $x_i > 0$, $c > 0$;

$$\pi_{\text{PRIOR}}(x) := \exp\left(-\alpha \sum_{i=1}^n x_i \log \frac{x_i}{c}\right)$$

discontinuities

strategy: consider finite difference approximation of derivative and assume it follows an impulse noise probability distribution

let $x : [0,1] \to \mathbb{R}$, $x_i := x(s_i)$, $s_i = ih$, $h = 1/n$, $i = 1, \dots, n$;

$$\pi_{\text{PRIOR}}(x) := \left(\frac{\alpha}{\pi}\right)^n \prod_{i=1}^n (1 + \alpha^2 (x_i - x_{i-1})^2)^{-1}$$

Cauchy density with finite difference approximation

Markov random fields

let X be an $\mathbb{R}^n$-valued random variable

neighborhood system related to X is a collection of index sets

$$N = \{ N_i : i = 1, \dots, n \}, \quad N_i \subset \{ 1, \dots, n \}$$

with the following properties

(i) $i \notin N_i$

(ii) $i \in N_j$ iff $j \in N_i$

the set N_i is the index set of the neighbors of the i-th component of X

we say that X is a discrete MRF with respect to the neighborhood system N if

$$\pi_{X_i}(\tilde{x} \mid x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \pi_{X_i}(\tilde{x} \mid x_j, \ j \in N_i)$$

Hammersley-Clifford theorem states that probability density of an MRF is of form $\pi(x) \propto \exp\left(-\sum_{i=1}^{n} V_i(x)\right)$

total variation density:

let $u : D \to \mathbb{R}$, $u \in L^1(D)$, $D \subset \mathbb{R}^n$

total variation of u

$$TV(u) := \sup \left\{ \int_D u \nabla \cdot \phi \, ds : \phi \in C_0^1(D, \mathbb{R}^n), \|\phi\|_{L^\infty(D)} \right\}$$

test function

a function is said to have bounded variation if

$$TV(u) < \infty$$

let $x : [0,1] \to \mathbb{R}$, $x_i := x(s_i)$, $h_{ij} = |s_i - s_j|$, $i, j = 1, \dots, n$, $i \neq j$

$$\pi_{\text{PRIOR}}(x) \propto \exp\left(-\alpha \sum_{i=1}^{n} V_i(x)\right) = \exp\left(-\alpha \sum_{j \in N_i} h_{ij} |x_i - x_j|\right)$$

sample-based densities

suppose $\pi(x)$ is the probability density of a random variable X, and we have a large number of realizations of X, $\{x_i\}_{i=1}^n$;
the objective is to approximate π based on $\{x_i\}_{i=1}^n$

example: assume we can measure n realizations $\{x_i\}_{i=1}^n$ of a random variable X

- non-parametric approach: compute a histogram based on $\{x_i\}_{i=1}^n$ and infer what underlying distribution is
- parametric approach: propose a parametric model; compute maximum likelihood estimate

$X \sim \mathcal{N}(\bar{x}, \sigma^2)$; parameters to be estimated: $\theta = \begin{pmatrix} \bar{x} \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$

likelihood function:

$$\begin{aligned}
\pi_{\text{LIKE}}(\{x_i\}_{i=1}^n | \theta) &= \prod_{i=1}^{n} \pi(x_i | \theta) \\
&= (2\pi\theta_2)^{-n/2} \exp\left(-(2\theta_2)^{-1} \sum_{i=1}^{n} (x_i - \theta_1)^2\right) \\
&= \exp\left(-(2\theta_2)^{-1} \sum_{i=1}^{n} (x_i - \theta_1)^2 - \frac{n}{2} \log(2\pi\theta_2)\right) \\
&= \exp\left(-g(\{x_i\}_{i=1}^n | \theta)\right)
\end{aligned}$$

$$d_\theta g(\{x_i\}_{i=1}^n | \theta) = \begin{pmatrix} -\theta_2^{-1} \sum_{i=1}^{n} x_i - \frac{n}{\theta_2} \theta_1 \\ -(2\theta_2^2)^{-1} \sum_{i=1}^{n} (x_i - \theta_1)^2 + \frac{n}{2\theta_2} \end{pmatrix}$$

$$\Rightarrow \quad \bar{x}_{OPT} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \sigma^2_{OPT} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{OPT})^2$$

- the unknown X is a random variable whose probability distribution – called the prior distribution – is denoted by $\pi_{PRIOR}(x)$
- assuming that a Gaussian prior is justifiable, we use the parametric model

$$\pi_{PRIOR}(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-(2\sigma^2)^{-1}(x-\bar{x})^2\right)$$

Gaussian densities

due to the central limit theorem, Gaussian densities are often very good approximations to non-Gaussian distributions when the observation is based on a large number of mutually independent random events

Gaussian n-variate random variable:

[def]

let $\bar{x} \in \mathbb{R}^n$, $\Gamma > 0$; a Gaussian n-variate random variable X with mean $\bar{x}$ and covariance Γ is a random variable with the probability density

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

we use the notation

$$X \sim \mathcal{N}(\bar{x}, \Gamma)$$

[def]

Gaussian random variables are often defined through the Fourier transform (or characteristic function): a random variable X is Gaussian if

$$\mathbb{E}[\exp(-i\xi^T X)] = \exp(-i\xi^T \bar{x} - \frac{1}{2}\xi^T \Gamma \xi)$$

with $\bar{x} \in \mathbb{R}^n$, $\Gamma \geq 0$

[lem]

let $\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix} > 0$, $\Gamma_{11} \in \mathbb{R}^{k,k}$, $\Gamma_{22} \in \mathbb{R}^{n-k,n-k}$, $k<n$, $\Gamma_{12} = \Gamma_{21}^T$;

then, the Schur complements

$$\tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21} \quad \text{and} \quad \tilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}$$

are invertible matrices and

$$\Gamma^{-1} = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \tilde{\Gamma}_{11}^{-1} \end{pmatrix} \quad (*)$$

proof:

$$|\Gamma| = \begin{vmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{vmatrix} = \begin{vmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12} \end{vmatrix} = \begin{vmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \tilde{\Gamma}_{11} \end{vmatrix} = |\Gamma_{11}||\tilde{\Gamma}_{11}| \neq 0$$

$$\Rightarrow |\tilde{\Gamma}_{11}| \neq 0$$

similarly, we can prove that $\tilde{\Gamma}_{22}$ is invertible;

the proof of $(*)$ follows from Gaussian elimination;

consider the linear system

$$\begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

by eliminating x_2 from the second equation we get

$$x_2 = \Gamma_{22}^{-1}(y_2 - \Gamma_{21}x_1)$$

substituting x_2 into the first equation we obtain

$$(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21})\, x_1 = y_1 - \Gamma_{12}\Gamma_{22}^{-1} y_2$$

$$\Leftrightarrow \qquad x_1 = \tilde{\Gamma}_{22}^{-1} y_1 - \tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} y_2$$

similarly, we solve 2nd equation for x_2;

claim follows

(*) $\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} = \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}$ by symmetry of Γ^{-1}

[thm] Let $X : \Omega \to \mathbb{R}^n$, $Y : \Omega \to \mathbb{R}^k$ be two Gaussian random variables whose joint probability density $\pi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}_+$ is of the form

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}\begin{pmatrix} x-\bar{x} \\ y-\bar{y}\end{pmatrix}^T \begin{pmatrix}\Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22}\end{pmatrix}^{-1} \begin{pmatrix} x-\bar{x} \\ y-\bar{y}\end{pmatrix}\right) \qquad (*)$$

with $\Gamma = \begin{pmatrix}\Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22}\end{pmatrix} > 0$, $\Gamma_{11} \in \mathbb{R}^{n,n}$, $\Gamma_{22} \in \mathbb{R}^{k,k}$, $\Gamma_{12} = \Gamma_{21}^T$; then the probability distribution of X conditioned on $Y = y$, $\pi(\cdot|y) : \mathbb{R}^n \to \mathbb{R}_+$ is of the form

$$\pi(x|y) \propto \exp\left(-\frac{1}{2}(x-\tilde{x})^T \tilde{\Gamma}_{22}^{-1}(x-\tilde{x})\right)$$

where $\tilde{x} = \bar{x} + \Gamma_{12}\Gamma_{22}^{-1}(y-\bar{y})$ and $\tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$ is the Schur complement of Γ

proof: WLOG, let $\bar{x} = \bar{y} = 0$; by Bayes' formula, we have $\pi(x|y) \propto \pi(x,y)$; with (*) and (*)

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}\begin{pmatrix} x-\bar{x} \\ y-\bar{y}\end{pmatrix}^T \begin{pmatrix}\Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22}\end{pmatrix}^{-1} \begin{pmatrix} x-\bar{x} \\ y-\bar{y}\end{pmatrix}\right) \qquad \bar{x} = \bar{y} = 0$$

$$\overset{(*)}{=} \exp\left(-\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \tilde{\Gamma}_{11}^{-1} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right)$$

$$\overset{(*)}{=} \exp\left(-\frac{1}{2}\left(x^T\tilde{\Gamma}_{22}^{-1}x - 2x^T\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}y + y^T\tilde{\Gamma}_{11}^{-1}y\right)\right)$$

by completing the quadratic form in the exponential into squares, we can express the joint distribution as

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}(x-\Gamma_{12}\Gamma_{22}^{-1}y)^T\tilde{\Gamma}_{22}^{-1}(x-\Gamma_{12}\Gamma_{22}^{-1}y) + c\right)$$

where $c = y^T\left(\tilde{\Gamma}_{11}^{-1} - \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}\right)y$

[thm]

Let $X:\Omega\to\mathbb{R}^n$, $Y:\Omega\to\mathbb{R}^k$ be two Gaussian random variables whose joint probability density $\pi:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}_+$ is given by (*); then the marginal density of X is

$$\pi(x) = \int_{\mathbb{R}^k}\pi(x,y)\,dy \propto \exp\left(-\frac{1}{2}(x-\bar{x})^T\Gamma_{11}^{-1}(x-\bar{x})\right)$$

proof: WLOG, let $\bar{x}=\bar{y}=0$; let $L=\Gamma^{-1}$, and using the partitioning of L,

$$L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} = \begin{pmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1} & \tilde{\Gamma}_{11}^{-1} \end{pmatrix},$$

we have

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}\left(x^TL_{11}x + 2x^TL_{12}y + y^TL_{22}y\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}(y+L_{22}^{-1}L_{21}x)^TL_{22}(y+L_{22}^{-1}L_{21}x) + x^T(L_{11}-L_{12}L_{22}^{-1}L_{21})x\right)$$

consequently,

$$\pi(x) = \int_{\mathbb{R}^k} \pi(x,y)\,dy \propto \exp\left(-\frac{1}{2}\, x^T \tilde{L}_{22} x\right)$$

by (*) the Schur complement of L_{22} is the first block of the inverse of L, i.e., $\tilde{L}_{22}^{-1} = (L^{-1})_{11} = \Gamma_{11}$

Linear inverse problems

Let $Y = KX + N$, $K \in \mathbb{R}^{m,n}$, with random variables $X : \Omega \to \mathbb{R}^n$, $Y, N : \Omega \to \mathbb{R}^m$; assume further that X and N are mutually independent Gaussian random variables with probability densities

$$\pi_{\text{PRIOR}}(x) \propto \exp\left(-\frac{1}{2}(x-\bar{x})^T \Gamma_{\text{PRIOR}}^{-1}(x-\bar{x})\right)$$

and $$\pi_{\text{NOISE}}(\eta) \propto \exp\left(-\frac{1}{2}(\eta-\bar{\eta})^T \Gamma_{\text{NOISE}}^{-1}(\eta-\bar{\eta})\right)$$

$$\pi_{\text{POST}}(x|y) = \pi_{\text{PRIOR}}(x)\,\pi_{\text{NOISE}}(y-Kx)$$
$$\propto \exp\left(-\frac{1}{2}(x-\bar{x})^T \Gamma_{\text{PRIOR}}^{-1}(x-\bar{x}) - \frac{1}{2}(y-Kx-\bar{\eta})^T \Gamma_{\text{NOISE}}^{-1}(y-Kx-\bar{\eta})\right)$$

Since X and N are Gaussian, Y is also Gaussian; we have

$$\mathbb{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}, \qquad \bar{y} = K\bar{x} - \bar{\eta}$$

since $\mathbb{E}[(X-\bar{x})(X-\bar{x})^T] = \Gamma_{\text{PRIOR}}$, we have

$$\mathbb{E}[(Y-\bar{y})(Y-\bar{y})^T] = \mathbb{E}[(K(X-\bar{x}) + (N-\bar{\eta}))(K(X-\bar{x}) + (N-\bar{\eta}))^T]$$
$$= K\Gamma_{\text{PRIOR}}K^T + \Gamma_{\text{NOISE}}$$

moreover

$$\mathbb{E}[(X-\bar{x})(Y-\bar{y})^T] = \mathbb{E}[(X-\bar{x})(K(X-\bar{x}) + (N-\bar{\eta}))^T] = \Gamma_{\text{PRIOR}}K^T$$

So, overall we have

$$\mathrm{Cov}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\right] = \mathbb{E}\left[\begin{pmatrix} X-\bar{x} \\ Y-\bar{y} \end{pmatrix}\begin{pmatrix} X-\bar{x} \\ Y-\bar{y} \end{pmatrix}^T\right]$$

$$= \begin{pmatrix} \Gamma_{\text{PRIOR}} & \Gamma_{\text{PRIOR}} K^T \\ K\Gamma_{\text{PRIOR}} & K\Gamma_{\text{PRIOR}} K^T + \Gamma_{\text{NOISE}} \end{pmatrix}$$

Hence, the joint probability density of X and Y is of the form

$$\pi(x,y) \propto \exp\left(-\frac{1}{2}\begin{pmatrix} x-\bar{x} \\ y-\bar{y} \end{pmatrix}^T \begin{pmatrix} \Gamma_{\text{PRIOR}} & \Gamma_{\text{PRIOR}} K^T \\ K\Gamma_{\text{PRIOR}} & K\Gamma_{\text{PRIOR}} K^T + \Gamma_{\text{NOISE}} \end{pmatrix}^{-1} \begin{pmatrix} x-\bar{x} \\ y-\bar{y} \end{pmatrix}\right)$$

[thm]

assume that $X : \Omega \to \mathbb{R}^n$ and $N : \Omega \to \mathbb{R}^m$ are mutually independent Gaussian random variables,

$$X \sim \mathcal{N}(\bar{x}, \Gamma_{\text{PRIOR}}), \qquad N \sim \mathcal{N}(\bar{\eta}, \Gamma_{\text{NOISE}}),$$

with $\Gamma_{\text{PRIOR}} \in \mathbb{R}^{n,n}$, $\Gamma_{\text{PRIOR}} > 0$, $\Gamma_{\text{NOISE}} \in \mathbb{R}^{m,m}$, $\Gamma_{\text{NOISE}} > 0$; moreover, assume $Y = KX + N$, with $K \in \mathbb{R}^{m,n}$ known; then the posterior density of X given the measurement $Y = y$ is

$$\pi_{\text{POST}}(x|y) \propto \exp\left(-\frac{1}{2}(x-\tilde{x})^T \Gamma_{\text{POST}}^{-1}(x-\tilde{x})\right)$$

where $\tilde{x} = \bar{x} + \Gamma_{\text{PRIOR}} K^T (K\Gamma_{\text{PRIOR}} K^T + \Gamma_{\text{NOISE}})^{-1}(y - K\bar{x} - \bar{\eta})$

↳ posterior mean

and $\Gamma_{\text{POST}} = \Gamma_{\text{PRIOR}} - \Gamma_{\text{PRIOR}} K^T (K\Gamma_{\text{PRIOR}} K^T + \Gamma_{\text{NOISE}})^{-1} K\Gamma_{\text{PRIOR}}$ (*)

↳ posterior covariance

the posterior can be directly derived from Bayes' formula for $\pi_{\text{POST}}(x|y)$ by arranging the quadratic form of the exponent according to the degree of x such a procedure gives the following alternative ;

representation

posterior covariance: $\Gamma_{\text{POST}} = (\Gamma_{\text{PRIOR}}^{-1} + K^T \Gamma_{\text{NOISE}}^{-1} K)^{-1}$

posterior mean: $\tilde{x} = (\Gamma_{\text{PRIOR}}^{-1} + K^T \Gamma_{\text{NOISE}}^{-1} K)^{-1} (K^T \Gamma_{\text{NOISE}}^{-1} (y - \bar{\eta}) + \Gamma_{\text{PRIOR}}^{-1} \bar{x})$

in the purely Gaussian case, the center point $\tilde{x}$ is simultaneously the maximum a posteriori estimator and the conditional mean:

$$\tilde{x} = x_{\text{MAP}} = x_{\text{CM}}$$

where

$$x_{\text{MAP}} = \operatorname{argmax} \pi_{\text{POST}}(x|y)$$

and

$$x_{\text{CM}} = \int x \, \pi_{\text{POST}}(x|y) \, dy$$

similarly, Γ_{POST} is the conditional covariance

example: let $X \sim \mathcal{N}(0, \gamma^2 I)$; we refer to this prior as Gaussian white noise prior; moreover, we assume that the noise is white noise: $N \sim \mathcal{N}(0, \sigma^2 I)$; we have

$$\tilde{x} = \gamma^2 K^T (\gamma^2 K K^T + \sigma^2 I)^{-1} y = K^T (K K^T + \alpha I)^{-1} y$$

where $\alpha = \sigma^2 / \gamma^2$

$$\tilde{x} = K^T (K K^T + \alpha I)^{-1} y = (K^T K + \alpha I)^{-1} K^T y$$

$$\pi_{\text{POST}}(x|y) \propto \exp(-\varphi(x|y))$$

in particular, $\varphi(x|y) = (2\gamma^2)^{-1} \|x\|^2 + (2\sigma^2)^{-1} \|y - Kx\|^2$

now, suppose $X \sim \mathcal{N}(0, \gamma^2 \Gamma_{\text{PRIOR}})$; then

$$\pi_{\text{PRIOR}}(x) \propto \exp(-(2\gamma^2)^{-1} x^T \Gamma_{\text{PRIOR}}^{-1} x)$$

$$\pi_{\text{LIKE}}(y|x) \propto \exp(-(2\sigma^2)^{-1} \|y - Kx\|^2)$$

by Bayes' formula

$$\pi_{POST}(x|y) \propto \pi_{LIKE}(y|x)\,\pi_{PRIOR}(x)$$
$$\propto \exp\left(-(2\sigma^2)^{-1}\|y-Kx\|^2-(2\gamma^2)^{-1}x^T\Gamma_{PRIOR}^{-1}x\right)$$
$$= \exp(-\varphi(x|y))$$

where $\varphi(x|y) = (2\sigma^2)^{-1}\|y-Kx\|^2+(2\gamma^2)^{-1}x^T\Gamma_{PRIOR}^{-1}x$

since $\Gamma_{PRIOR}>0$, $\Gamma_{PRIOR}^{-1} = L^TL$

consequently, $x^T\Gamma_{PRIOR}^{-1}x = x^TL^TLx = \|Lx\|^2$

$$\varphi(x) := 2\sigma^2\varphi(x|y) = \|y-Kx\|^2+\alpha\|Lx\|^2,\quad \alpha=\sigma/\gamma$$

Gaussian smoothness priors

Tikhonov (smoothness) regularization:

$$reg:\mathbb{R}^n\to\mathbb{R},\qquad reg(x) = \frac{\alpha}{2}\|Lx\|^2$$

let $L:\mathbb{R}^n\to\mathbb{R}^m$ denote a discrete approximation of Laplacian in $\mathbb{R}^n$

let $x:[0,1]\to\mathbb{R}$, $x_i := x(s_i)$, $s_i = ih$, $h=1/n$, $i=1,\dots,n$;

then, e.g.,

$$L = h^{-2}\begin{pmatrix} -1 & 2 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & 2 & -1 \end{pmatrix}$$

smoothness prior:

$$\pi_{PRIOR}(x) \propto \exp\left(-\frac{1}{2}\|L(x-\bar{x})\|^2\right) = \exp\left(-\frac{1}{2}(x-\bar{x})^TL^TL(x-\bar{x})\right)$$

we seek interpretation through a limiting process

we will assume $\bar{x}=0$

a random variable $W\in\mathbb{R}^m$ is called <u>pure or orthonormal white noise</u> if $W\sim\mathcal{N}(0,I_m)$

assume $X \in \mathbb{R}^n$ is a Gaussian zero mean random variable;
the matrix $L \in \mathbb{R}^{m,n}$ is called a <u>whitening matrix</u> if

$$LX = W \in \mathbb{R}^m$$

is pure white noise

assume $X \in \mathbb{R}^n$ is Gaussian with covariance $\Gamma > 0$,
$\Gamma = CC^T$, $C > 0$ $(C \in \mathbb{R}^{n,n})$
then the variable $Y = C^{-1}X$ is pure white noise
indeed $\mathbb{E}[YY^T] = \mathbb{E}[C^{-1}XX^T(C^{-1})^T]$
by linearity of $\mathbb{E}$

$$\mathbb{E}[C^{-1}XX^T(C^{-1})^T] = C^{-1}\,\mathbb{E}[XX^T]\,(C^{-1})^T$$

since $\Gamma = \mathbb{V}ar[X] = \mathbb{E}[XX^T] - \bar{x}\bar{x}^T$, $\bar{x} = 0$, and $\Gamma = CC^T$
we have $C^{-1}\,\mathbb{E}[XX^T]\,(C^{-1})^T = C^{-1}\Gamma(C^{-1})^T = C^{-1}CC^T(C^{-1})^T = I_n$
so, overall $\mathbb{E}[YY^T] = I_n$
Hence, the inverse of the Cholesky factor C of
the covariance matrix Γ provides a natural
whitening matrix of the random variable X
conversely, assume a $L \in \mathbb{R}^{m,n}$ is given; our goal is
to construct a random variable X such that
L is <u>almost</u> a whitening matrix of X
let $L = USV^T \in \mathbb{R}^{m,n}$ be given, with

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p > \sigma_{p+1} = \dots = \sigma_q = 0, \qquad q = \min(m,n)$$

$$\ker L = \operatorname{span}\{v_{p+1}, \dots, v_q\} \subseteq \mathbb{R}^n$$

[lem]

let $L = USV^T \in \mathbb{R}^{m,n}$, $Q = (v_{p+1} \dots v_q) \in \mathbb{R}^{n,q-p}$, and let $W \in \mathbb{R}^m$

and $\tilde{W} \in \mathbb{R}^m$ be two independent white noise random variables, and let

$$X = L^\dagger W + \gamma Q \tilde{W}$$

with pseudo-inverse $L^\dagger$, and $\gamma > 0$ arbitrary; then X has covariance $\Gamma = L^\dagger (L^\dagger)^T + \gamma^2 QQ^T$, $\Gamma^{-1} = L^T L + \gamma^{-2} QQ^T$

proof: by mutual independence of W and $\tilde{W}$, we have

$$\mathbb{C}ov[X,X] = \mathbb{E}[XX^T] = L^\dagger \mathbb{E}[WW^T](L^\dagger)^T + \gamma^2 Q \mathbb{E}[\tilde{W}\tilde{W}^T] Q^T$$
$$= L^\dagger (L^\dagger)^T + \gamma^2 QQ^T = \Gamma$$

this matrix is invertible:

for $j \le p$: $\Gamma v_j = L^\dagger (L^\dagger)^T v_j + \gamma^2 QQ^T v_j = L^\dagger (L^\dagger)^T v_j = \frac{1}{\sigma_j^2}$

for $j > p$: $\Gamma v_j = L^\dagger (L^\dagger)^T v_j + \gamma^2 QQ^T v_j = \gamma^2 QQ^T v_j = \gamma^2 v_j$

$$\Gamma^{-1} = \sum_{i=1}^{p} \sigma_i^2 v_i v_i^T + \frac{1}{\gamma^2} \sum_{i=p+1}^{n} v_i v_i^T = L^T L + \gamma^{-2} QQ^T$$

let $X = L^\dagger W + \gamma Q \tilde{W}$; then

$$LX = LL^\dagger W + \gamma L Q \tilde{W} = LL^\dagger W = UU^T W$$

$LL^\dagger$: orthogonal projection on range of L

Q: spans null space

if $\ker L$ is nontrivial, the smoothness prior is not a proper probability density

$$\pi_{\text{PRIOR}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2\right) = \exp\left(-\frac{1}{2\gamma^2} \sum_{i=1}^{p} \sigma_i^2 (v_i^T x)^2\right)$$

by setting $V = \text{span}\{v_1, \ldots, v_p\}$ we have

$$\int_V \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2\right) dx = \int_V \exp\left(-\frac{1}{2\gamma^2} \sum_{i=1}^{p} \sigma_i^2 (v_i^T x)^2\right) dx$$

$$= (2\pi)^{p/2}\, \gamma^p \left(\prod_{i=1}^{n} \sigma_i\right)^{-1} < \infty$$

however, for $n > p$, $\int_V \pi_{\text{PRIOR}}(x)\,dx = \infty$

a prior density with nonintegrability property is called an improper density

interpreting the posterior distribution

example: inverse problem: determine x given $y = x + \eta$

statistical model:

$$Y = X + N, \quad X \sim \mathcal{N}(0,1), \quad N \sim \mathcal{N}(0,\sigma^2)$$

consequently,

$$\pi_{\text{POST}}(x|y) \propto \exp\left(-\tfrac{1}{2}x^2 - (2\sigma^2)^{-1}(y-x)^2\right)$$

$$\pi_{\text{POST}}(x|y) \propto \exp\left(-(1+\sigma^2)(2\sigma^2)^{-1}\left(x - (1+\sigma^2)^{-1}y\right)^2\right)$$

$$x_{CM} = (1+\sigma^2)^{-1}y \quad \text{and} \quad \gamma^2 = \sigma^2(1+\sigma^2)^{-1}$$

in this case it is easy to calculate the credibility interval; for example, the 90% interval is

$$ci(90) \approx [x_{CM} - c\gamma,\ x_{CM} + c\gamma], \quad c \approx 1.64$$

assume data corresponds to a "true" value of x, say $x = x_{\text{TRUE}} > 0$ and that noise is negligible, i.e., $y \approx x_{\text{TRUE}} > 0$

it may happen that in this case

$$x_{\text{TRUE}} \notin ci(90)$$

in fact, if $y\gamma > c$, then

$$x_{CM} + c\gamma < x_{CM} + y\gamma^2 = y \approx x_{TRUE}$$

conclusion $x_{TRUE} \approx y > c/\gamma$ is rather improbable

$$\int_{c/\gamma}^{\infty} \pi_{PRIOR}(x)\,dx < \int_{c}^{\infty} \pi_{PRIOR}(x)\,dx < 0.5$$