19. Conjugate Priors

ECE 830, Spring 2014

Choosing a Prior

Two approaches:

1. Informative (or "subjective") priors:
   - design/choose priors that are compatible with prior knowledge of unknown parameters
   - can be inpractical in complicated problems with many parameters
   - injecting subjective opinion into analysis contrary to making scientific analysis as objective as possible.
2. Non-informative priors:
   - attempt to remove subjectiveness from Bayesian procedures
   - designs are often based on invariance arguments

Selecting an Informative Prior

Clearly, the most important objective is to choose the prior $p(\theta)$ that best reflects the prior knowledge available to us.

In general, however, our prior knowledge is imprecise and any number of prior densities may aptly capture this information.

Moreover, usually the optimal estimator can't be obtained in closed-form.

Therefore, sometimes it is desirable to choose a prior density that models prior knowledge and is nicely matched in functional form to $p(x|\theta)$ so that the optimal estimator (and posterior density) can be expressed in a simple fashion.

Conjugate Priors

Idea: Given $p(x|\theta)$, choose $p(\theta)$ so that $p(\theta|x) \propto p(x|\theta)\, p(\theta)$ has a simple functional form.

Conjugate priors: choose $p(\theta) \in \mathcal{P}$, where $\mathcal{P}$ is a family of densities (e.g., Gaussian family) so that the posterior density also belongs to that family.

Definition

$p(\theta)$ is a *conjugate prior* for $p(x|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

Example: Conjugate priors for exponential random variables

$$X_1, \ldots, X_N \sim \text{exponential}(\theta)$$

$$p(x|\theta) = \prod_{n=1}^{N} \theta e^{-\theta x_n} = \theta^N e^{-\theta t}$$

where $t := \sum x_n$.

Example: (cont.)

Let $\theta \sim \text{Gamma}(\alpha, \beta)$, so that $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$ for $\theta \in [0, \infty)$. Then

$$p(x, \theta) =$$

$$p(x) =$$

$$=$$

$$=$$

$$p(\theta|x) = \frac{p(x, \theta)}{p(x)}$$

$$=$$

$$=$$

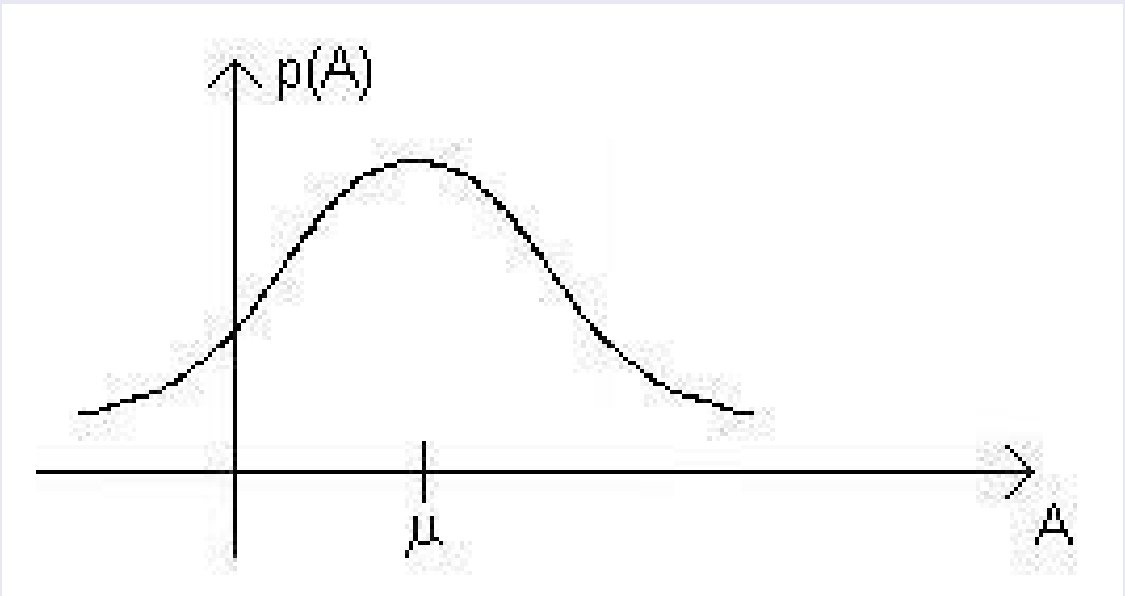
Thus the Gamma prior is conjugate for the exponential distribution!

Example: DC Level in AWGN

$$x_n = A + w_n \ , \ n = 1, \cdots, N; \qquad w_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Rather than modeling $A \sim \text{Uniform}(-A_0, A_0)$ (which did not yield a closed-form estimator) consider

$$p(A) = \frac{1}{\sqrt{2\pi {\sigma_A}^2}} \exp\left\{ -\frac{1}{2{\sigma_A}^2}(A - \mu)^2 \right\}$$



Example: (cont.)

With $\mu = 0$ and $\sigma_A = \frac{1}{3}A_0$ this Gaussian prior also reflects prior knowledge that it is unlikely for $|A| \geq A_0$.
The Gaussian prior is also conjugate to the Gaussian likelihood

$$p(x|A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N}(x_n - A)^2\right]$$

so that the resulting posterior density is also a simple Gaussian, as shown next. First note that

$$p(x|A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} x_n^2\right] \cdot \exp\left[-\frac{1}{2\sigma^2}\left(NA^2 - 2NA\bar{x}\right)\right]$$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.

Example: (cont.)

$$
\begin{aligned}
p(A|x) &= \frac{p(x|A)\, p(A)}{\int p(x|a)\, p(a)\, da} \\
&= \frac{\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(NA^2 - 2NA\bar{x}) + \frac{1}{{\sigma_A}^2}(A-\mu)^2\right)\right]}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(Na^2 - 2Na\bar{x}) + \frac{1}{{\sigma_A}^2}(a-\mu)^2\right)\right] da} \\
&= \frac{e^{-\frac{1}{2}Q(A)}}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}Q(a)} da}
\end{aligned}
$$

where

$$
Q(A) = \frac{N}{\sigma^2}A^2 - \frac{2NA\bar{x}}{\sigma^2} + \frac{A^2}{{\sigma_A}^2} - \frac{2\mu A}{{\sigma_A}^2} + \frac{\mu^2}{{\sigma_A}^2}
$$

Now let

$$
\begin{aligned}
\sigma^2_{A|X} &\equiv \frac{1}{\frac{N}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\
\mu^2_{A|X} &\equiv \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu}{{\sigma_A}^2}\right)\sigma^2_{A|X}
\end{aligned}
$$

Example: (cont.)

Then by "completing the square" we have

$$\begin{aligned} Q(A) &= \frac{1}{\sigma^2_{A|X}}\left(A^2 - 2\mu_{A|X}A + \mu^2_{A|X}\right) - \frac{\mu^2_{A|X}}{\sigma^2_{A|X}} + \frac{\mu^2}{{\sigma_A}^2} \\ &= \frac{1}{\sigma^2_{A|X}}\left(A^2 - \mu_{A|X}\right)^2 - \frac{\mu^2_{A|X}}{\sigma^2_{A|X}} + \frac{\mu^2}{{\sigma_A}^2} \end{aligned}$$

Hence

$$p(A|x) = \frac{\exp\left[-\frac{1}{2\sigma^2_{A|X}}(A-\mu_{A|X})^2\right]\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|X}}{\sigma^2_{A|X}}\right)\right]}{\int_{-\infty}^{\infty} \underbrace{\exp\left[-\frac{1}{2\sigma^2_{A|X}}(a-\mu_{A|X})^2\right]}_{\text{"unnormalized" Gaussian density}} \underbrace{\exp\left[-\frac{1}{2}\left(\frac{\mu^2}{{\sigma_A}^2} - \frac{\mu^2_{A|X}}{\sigma^2_{A|X}}\right)\right]}_{\text{Constant, indep. of } a} da}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2_{A|X}}}\exp\left[-\frac{1}{2\sigma^2_{A|X}}(A-\mu_{A|X})^2\right]$$

$$A|x \sim \mathcal{N}(\mu_{A|X}, \sigma^2_{A|X})$$

Example: (cont.)

Now

$$\begin{aligned}\widehat{A} &= \mathbb{E}[A|x] = \mu_{A|X} = \frac{\frac{N}{\sigma^2}\bar{x} + \frac{\mu}{{\sigma_A}^2}}{\frac{N}{\sigma^2} + \frac{1}{{\sigma_A}^2}} \\ &= \left(\frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/N}\right)\bar{x} + \left(\frac{\sigma^2/N}{{\sigma_A}^2 + \sigma^2/N}\right)\mu \\ &= \alpha\bar{x} + (1-\alpha)\mu\end{aligned}$$

where

$$0 < \alpha = \frac{{\sigma_A}^2}{{\sigma_A}^2 + \sigma^2/N} < 1$$

Interpretation

1. When there is little data ($\sigma_A^2 \ll \frac{\sigma^2}{N}$), α is small and $\widehat{A} \approx \mu$.
2. When there is a lot of data ($\sigma_A^2 \gg \frac{\sigma^2}{N}$), $\alpha \approx 1$ and $\widehat{A} \approx \bar{x}$.

Interplay between data and prior knowledge

The Multivariate Gaussian Model

The multivariate Gaussian model is the most important Bayesian tool in signal processing. It leads directly to the celebrated Wiener and Kalman filters.
Consider the Bayesian statistical model

$$x = H\theta + W$$

where

θ	is	unknown, $p \times 1$
H	is	known, $N \times p$
θ	$\sim$	$\mathcal{N}(\mu_\theta, R_\theta)$
W	$\sim$	$\mathcal{N}(0, R_W)$
θ and W	are	independent
$R_\theta, R_W, \mu_\theta$	are	known

This model amounts to a signal subspace with a Gaussian prior on θ and a Gaussian conditional distribution of x given θ.

Theorem

The posterior distribution of $\theta|x$ is

$$\theta|x \sim \mathcal{N}(\mu_{\theta|X}, R_{\theta|X})$$

where

$$\begin{aligned}
\mu_{\theta|X} &= \mu_\theta + R_\theta H^\top \left(H R_\theta H^\top + R_W\right)^{-1} (x - H\mu_\theta) \\
&= \mu_\theta + \left(H^\top R_W^{-1} H + R_\theta^{-1}\right)^{-1} H^\top R_W^{-1} (x - H\mu_\theta) \\
R_{\theta|X} &= R_\theta - R_\theta H^\top \left(H R_\theta H^\top + R_W\right)^{-1} H R_\theta \\
&= \left(H^\top R_W^{-1} H + R_\theta^{-1}\right)^{-1}
\end{aligned}$$

Proof

First, note the x and θ are jointly Gaussian:

$$\begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} H & I_p \\ I_p & 0 \end{bmatrix} \begin{bmatrix} \theta \\ W \end{bmatrix}.$$

Since

$$\begin{bmatrix} \theta \\ W \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_\theta \\ 0 \end{bmatrix}, \begin{bmatrix} R_\theta & 0 \\ 0 & R_W \end{bmatrix} \right),$$

we have

$$\begin{bmatrix} x \\ \theta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} H\mu_\theta \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} HR_\theta H^\top + R_W & HR_\theta \\ R_\theta H^\top & R_\theta \end{bmatrix} \right).$$

Lemma

If

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \right),$$

then

$$Z_2 | Z_1 = z_1 \sim \mathcal{N}\left(\mu', R'\right)$$

where

$$\mu' := \mu_2 + R_{21} R_{11}^{-1} (z_1 - \mu_1)$$
$$R' := R_{22} - R_{21} R_{11}^{-1} R_{12}.$$

The theorem follows by applying this result to $\begin{bmatrix} x \\ \theta \end{bmatrix}$. The second version of each expression is a result of the following:

Matrix Inversion Lemma

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA.$$

Observations

- The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the optimal estimator (minimizing the Bayes risk) is

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(x) = \mu_{\theta|X} &= \mu_\theta + R_\theta H^\top (H R_\theta H^\top + R_W)^{-1}(x - H\mu_\theta) \\ &= \mu_\theta + (H^\top R_W^{-1} H + R_\theta^{-1})^{-1} H^\top R_W^{-1}(x - H\mu_\theta)\end{aligned}$$

  regardless of the loss function.
- $\widehat{\boldsymbol{\theta}}(x)$ is an affine function of x.
- $\widehat{\boldsymbol{\theta}}(x)$ is again multivariate Gaussian.
- Consider the case where $R_\theta = \sigma^2 I_p$ and $\sigma^2 \longrightarrow \infty$. This can be thought of as a "noncommittal" prior. Then $R_\theta^{-1} \longrightarrow 0_p$ and

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(x) = \mu_{\theta|X} &= \mu_\theta + (H^\top R_W^{-1} H)^{-1} H^\top R_W^{-1}(x - H\mu_\theta) \\ &=\end{aligned}$$

Example:

$x = s + w \;\;,\;\; w \sim \mathcal{N}(0, \sigma^2 I)$

$$
\begin{aligned}
p(s) &= \mathcal{N}(0, R_{ss}) \text{ indep. of } w \\
\mathbb{E}[x] &= \\
\mathbb{E}\left[xx^\top\right] &= \\
&= \\
\mathbb{E}\left[xs^\top\right] &= \mathbb{E}\left[ss^\top\right] + \mathbb{E}\left[ws^\top\right] \\
&= \\
\begin{bmatrix} x \\ s \end{bmatrix} &\sim \mathcal{N}
\end{aligned}
$$

Example: (cont.)

From our Bayesian perspective, we are interested in $p(s|x)$.

$$
\begin{aligned}
p(s|x) &= \frac{p(x,s)}{p(x)} \\
&= \frac{(2\pi)^{-N/2}\,(2\pi)^{-N/2}\,|R|^{-1/2}\,\exp\left\{-\frac{1}{2}[x^\top s^\top]R^{-1}\begin{bmatrix} x \\ s \end{bmatrix}\right\}}{(2\pi)^{-N/2}\,|R_{xx}|^{-1/2}\,\exp\left\{-\frac{1}{2}x^\top R_{xx}^{-1}x\right\}}
\end{aligned}
$$

In this formula we are faced with

$$
R^{-1} = \begin{bmatrix} R_{xx} & R_{xs} \\ R_{sx} & R_{ss} \end{bmatrix}^{-1}
$$

The inverse of this covariance matrix can be written as

$$
\begin{bmatrix} R_{xx} & R_{xs} \\ R_{sx} & R_{ss} \end{bmatrix}^{-1} = \begin{bmatrix} R_{xx}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -R_{xx}^{-1}R_{xs} \\ I \end{bmatrix} Q^{-1} \begin{bmatrix} -R_{sx}R_{xx}^{-1} & I \end{bmatrix}
$$

where $Q := R_{ss} - R_{sx}\,R_{xx}^{-1}\,R_{xs}$ is the Schur complement of R_{ss}.
(Verify this formula by applying RHS above to R to get I.)

Example: (cont.)

Furthermore,

$$\det R =$$

Substituting this expression into $p(s|x)$ we get

$$\begin{aligned} p(s|x) &= (2\pi)^{-N/2} \, |Q|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2}(s - R_{sx}R_{xx}^{-1}x)^\top Q^{-1}(s - R_{sx}R_{xx}^{-1}x) \right\} \\ s|x &\sim \mathcal{N} \end{aligned}$$

Thus the posterior mean of s is

$$\widehat{s} =$$

and the posterior variance is

Example: DC Level in AWGN

$$x_n = A + w_n,\ i = 1, \ldots, N$$

where A is an unknown scalar to be estimated and

$$\begin{aligned} A &\sim \mathcal{N}(\mu_A, {\sigma_A}^2) \\ w &\sim \mathcal{N}(0, {\sigma_w}^2)\ , \text{ indep. of y} \end{aligned}$$

This problem falls within the Gaussian linear model with

Example: (cont.)

Using the second formula for $\mu_{A|x}$, we obtain

$$\widehat{A}(x) = \mu_{A|x} = \mu_A + \left(\frac{1}{\sigma_w^2}\mathbf{1}^\top\mathbf{1} + \frac{1}{\sigma_A^2}\right)^{-1}\mathbf{1}^\top\frac{1}{\sigma_w^2}(x - \mathbf{1}\mu_A)$$

$$=$$

$$=$$

$$=$$

Example: (cont.)

In other words,

Thus

$$\widehat{A}(x) = (1-\alpha)\mu_A + \alpha\overline{x}$$

where

$$\alpha =$$

controls the tradeoff between prior knowledge and data.

Limiting cases:

$$\begin{aligned}
N \to \infty &\implies \alpha \to &&\implies \widehat{A} \to \\
N = 0 &\implies \alpha = &&\implies \widehat{A} = \\
\sigma_A^2 \to \infty &\implies \alpha \to &&\implies \widehat{A} \to \\
\sigma_A^2 = 0 &\implies \alpha = &&\implies \widehat{A} =
\end{aligned}$$

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in AWGN:

$$x = s + w$$

where x is the observed signal, s is the clean signal, and w is the noise. This can be modeled using a general linear model using $\theta = s$ and $H = I_N$. We can adopt a Gaussian prior for s:

$$S \sim \mathcal{N}(0, R_{ss}).$$

The Bayesian estimate of s is then

$$\widehat{s} = R_{ss} \left(R_{ss} + R_{ww}\right)^{-1} x.$$

Now suppose that R_{ss} and R_{ww} are simultaneously diagonalizable, meaning there exists an orthogonal matrix U such that

$$R_{ss} = U \Lambda_s U^\top$$

$$R_{ww} = U \Lambda_w U^\top$$

with Λ_s, Λ_w diagonal. For example, consider $R_{ww} = \sigma^2 I$ and R_{ss} arbitrary.

Then the estimator becomes

$$\begin{aligned}
\widehat{s} &= R_{ss}\left(R_{ss}+R_{ww}\right)^{-1}x \\
&= U\Lambda_s U^\top\left(U\Lambda_s U^\top + U\Lambda_w U^\top\right)^{-1}x \\
&= \\
&=
\end{aligned}$$

where

$$\Lambda = \begin{bmatrix}
\frac{\lambda_1^{(s)}}{\lambda_1^{(s)}+\lambda_1^{(w)}} & 0 & \cdots & 0 \\
0 & \frac{\lambda_2^{(s)}}{\lambda_2^{(s)}+\lambda_2^{(w)}} & \cdots & 0 \\
\vdots & \vdots & \ddots & \\
0 & 0 & \cdots & \frac{\lambda_N^{(s)}}{\lambda_N^{(s)}+\lambda_N^{(w)}}
\end{bmatrix}.$$

Interpretation:

- U is a change of basis matrix
- $s = U^\top x$ are coefficients of x in new basis
- $z = \Lambda s$ is a coordinate-wise rescaling of s
- $\widehat{s} = Uz$ is a reconstruction of s from z.

How should we interpret the weights

$$\lambda_i := \frac{\lambda_N^{(s)}}{\lambda_N^{(s)} + \lambda_N^{(w)}}?$$

Notice that

$$U^\top x = U^\top s + U^\top w$$
$$U^\top s \sim$$
$$U^\top w \sim$$

Writing

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix}$$

we have

$$u_i^\top S \sim \mathcal{N}(0, \lambda_i^s)$$
$$u_i^\top W \sim \mathcal{N}(0, \lambda_i^w)$$

Thus, λ_i reflects the proportion of the projection onto u_i that is due to the signal.

Example: Bandpass filtering

Suppose we observe

$$x = s + w$$

and we know *a priori* that the signal of interest occupies a certain passband.In other words, $|u_k^{\mathsf{H}} s|$ is large on average for certain DFT basis vectors u_k and small for others.

Example: (cont.)

How can we incorporate this prior knowledge into the prior for s? In other words, what should we use for R_{ss}?

Assume we can specify

$$\sigma_k^2 = \mathbb{E}\left[|u_k^{\mathsf{H}} S|^2\right],$$

the average signal energy at frequency k/N. Also assume that the signal content at different frequencies are independent. This amounts to assuming

$$U^{\mathsf{H}} S \sim \mathcal{N}(0, \Sigma)$$

as a prior, where

$$\Sigma = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}.$$

Example: (cont.)

Equivalently, the prior on S is

$$S \sim$$

Notice that the energy of S is

$$\begin{aligned}\mathbb{E}[S^{\mathsf{H}}S] &= \mathbb{E}[(U^{\mathsf{H}}S)^{\mathsf{H}}(U^{\mathsf{H}}S)] \\ &= \mathbb{E}[S^{\mathsf{H}}UU^{\mathsf{H}}S] = \sum_{k=0}^{N-1}\sigma_k^2.\end{aligned}$$

Thus, specifying σ_k^2 can be done as long as we know the total signal energy and the shape of the frequency response.
Assume the noise is iid:

$$R_{ww} = \sigma^2 I_N$$

with σ^2 known.

Example: (cont.)

Then the MMSE estimator is

$$\begin{aligned}\widehat{s} &= R_{ss}(R_{ss}+R_{ww})^{-1}x \\ &= U\Sigma U^{\mathsf{H}}(U(\Sigma+\sigma^2 I_N)U^{\mathsf{H}})^{-1}x \\ &= U[\Sigma(\Sigma+\sigma^2 I_N)^{-1}]U^{\mathsf{H}}x\end{aligned}$$

Notice that

$$\Sigma(\Sigma+\sigma^2 I_N)^{-1} = \begin{bmatrix} \frac{\sigma_0^2}{\sigma_0^2+\sigma^2} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_1^2}{\sigma_1^2+\sigma^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \frac{\sigma_{N-1}^2}{\sigma_{N-1}^2+\sigma^2} \end{bmatrix}.$$

Therefore, the MMSE estimator is a bandpass filter.

Example: (cont.)

Interpretation:

- $\sigma_K^2 \gg \sigma^2 \Longrightarrow$ keep most of signal
- $\sigma_K^2 \ll \sigma^2 \Longrightarrow$ kill most of signal
- $\sigma_K^2 \approx \sigma^2 \Longrightarrow$ keep some of signal