Information Form for the Gaussian, Likelihood Function, MAP Estimate and Regularized Least Squares, Gaussian Linear Models

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

August 31, 2017

Contents

- Chris Bishops' PRML book, Chapters 1 and 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapters 2 and 4

Information Form of the Gaussian

Information Form of the Gaussian

- As an alternative to representing a Gaussian in terms of its moments,

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

one can use natural/canonical parameters leading to the so called information form of the Gaussian.

- They are defined as:

$$\Lambda = \Sigma^{-1}, \boldsymbol{\xi} = \Sigma^{-1}\boldsymbol{\mu}$$

- They can be transformed back to moments as:

$$\boldsymbol{\mu} = \Lambda^{-1}\boldsymbol{\xi}, \Sigma = \Lambda^{-1}$$

- With direct substitution of the natural parameters in the moment parametrization, the information form of the Gaussian takes the following form:

$$\mathcal{N}_c(\boldsymbol{x} \mid \boldsymbol{\xi}, \Lambda) = (2\pi)^{-D/2} |\Lambda| \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^T \Lambda \boldsymbol{x} + \boldsymbol{\xi}^T \Lambda^{-1} \boldsymbol{\xi} - 2\boldsymbol{x}^T \boldsymbol{\xi}\right)\right\}$$

Information Form of the Gaussian

$$\mathcal{N}_c(\boldsymbol{x}\,|\,\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}\,|\boldsymbol{\Lambda}|\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{\xi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} - 2\boldsymbol{x}^T\boldsymbol{\xi}\right)\right\}$$

- One can easily derive the conditional of the multivariate Gaussian in information form (exponential family form)

$$p\left(\boldsymbol{x}_a\,|\,\boldsymbol{x}_b\right) = \mathcal{N}\left(\boldsymbol{x}_a\,|\,\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b), \boldsymbol{\Lambda}_{aa}^{-1}\right) \Rightarrow$$

$$p\left(\boldsymbol{x}_a\,|\,\boldsymbol{x}_b\right) = \mathcal{N}_c\left(\boldsymbol{x}_a\,|\,\boldsymbol{\Lambda}_{aa}\left(\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)\right), \boldsymbol{\Lambda}_{aa}\right)$$

$$= \mathcal{N}_c\left(\boldsymbol{x}_a\,|\,\underbrace{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b}_{\boldsymbol{\xi}_a} - \boldsymbol{\Lambda}_{ab}\boldsymbol{x}_b, \boldsymbol{\Lambda}_{aa}\right) \Rightarrow$$

$$p\left(\boldsymbol{x}_a\,|\,\boldsymbol{x}_b\right) = \mathcal{N}_c\left(\boldsymbol{x}_a\,|\,\boldsymbol{\xi}_a - \boldsymbol{\Lambda}_{ab}\boldsymbol{x}_b, \boldsymbol{\Lambda}_{aa}\right)$$

- This is a much easier form than that in terms of moments. That will not be the case for the marginal distributions as we see next.

Information Form of the Gaussian

$$\mathcal{N}_c(\boldsymbol{x}\,|\,\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}\,|\boldsymbol{\Lambda}|\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{\xi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} - 2\boldsymbol{x}^T\boldsymbol{\xi}\right)\right\}$$

- Similarly we can derive the marginal of the multivariate Gaussian in information form starting with $p(\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_b\,|\,\boldsymbol{\mu}_b,\boldsymbol{\Sigma}_{bb})$

- Utilizing an earlier result for $\boldsymbol{\Sigma}_{bb}$ and using the Woodbury formula

$$\boldsymbol{\Sigma}_{bb}^{-1} = \left(\boldsymbol{\Lambda}_{bb}^{-1} + \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1}\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\right)^{-1} = \boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}$$

- Thus: $p(\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_b\,|\,\boldsymbol{\mu}_b,\boldsymbol{\Sigma}_{bb}) \Rightarrow$

$$p(\boldsymbol{x}_b) = \mathcal{N}_c\left(\boldsymbol{x}_b\,|\left(\boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\right)\boldsymbol{\mu}_b, \boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\right)$$

$$= \mathcal{N}_c\left(\boldsymbol{x}_b\,|\,\underbrace{\left(\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b\right)}_{\boldsymbol{\xi}_b} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\underbrace{\left(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right)}_{\boldsymbol{\xi}_a}, \boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\right) \Rightarrow$$

$$p(\boldsymbol{x}_b) = \mathcal{N}_c\left(\boldsymbol{x}_b\,|\,\boldsymbol{\xi}_b - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\xi}_a, \boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\right)$$

Multiplication of Gaussians in Information Form

- Let us consider the multiplication of two Gaussian in information form. Expanding the product and keeping only x-terms, we obtain:

$$\mathcal{N}_c\left(\xi_1,\lambda_1\right)\mathcal{N}_c\left(\xi_2,\lambda_2\right)=\left(2\pi\right)^{-1/2}\sqrt{\lambda_1}\exp\left\{-\frac{1}{2}\left(x^2\lambda_1+\xi_1^2\lambda_1^{-1}-2x\xi_1\right)\right\}\times$$

$$\left(2\pi\right)^{-1/2}\sqrt{\lambda_2}\exp\left\{-\frac{1}{2}\left(x^2\lambda_2+\xi_2^2\lambda_2^{-1}-2x\xi_2\right)\right\}\propto$$

$$\exp\left\{-\frac{1}{2}\left(x^2\left(\lambda_1+\lambda_2\right)-2x\left(\xi_1+\xi_2\right)\right)\right\}\propto\mathcal{N}_c\left(\xi_1+\xi_2,\lambda_1+\lambda_2\right)$$

- This is much simpler than the moment based form:

$$\mathcal{N}\left(\mu_1,\sigma_1^2\right)\mathcal{N}\left(\mu_2,\sigma_2^2\right)\propto\exp\left\{-\frac{1}{2\sigma_1^2}\left(x^2-2x\mu_1\right)-\frac{1}{2\sigma_2^2}\left(x^2-2x\mu_2\right)\right\}\propto$$

$$\exp\left\{-\frac{1}{2}\left(x^2\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)-2x\left(\frac{\mu_1}{\sigma_1^2}+\frac{\mu_2}{\sigma_2^2}\right)\right)\right\}\propto\mathcal{N}\left(\frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2},\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$$

Likelihood Calculations (Additive & Multiplicative Noise Models), MAP Estimation and Regularized Least Squares

The Likelihood Function

- Consider Bayes' theorem

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}, \quad p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta)\, p(\theta) d\theta$$

- The quantity $p(\mathcal{D}|\theta)$ on the right-hand side of Bayes' theorem is evaluated for the observed data set $\mathcal{D}$ and can be viewed as a function of the parameter vector θ, in which case it is called the likelihood function.

- Given this definition of likelihood, we can state Bayes' theorem in words

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Frequentist Versus Bayesian Paradigms

- The likelihood $p(\mathcal{D}|\theta)$ is essential in both Bayesian and frequentist approaches but it is used in different roles.

- In a frequentist approach, θ is a fixed parameter computed by an estimator (e.g. maximum likelihood estimator). Error bars on this point estimate are computed by considering the distribution of all possible data sets $\mathcal{D}$ (e.g. variability of predictions between different bootstrap data sets)

- In the Bayesian approach, there is only one set of data $\mathcal{D}$ and the uncertainty in θ is introduced with appropriate prior and computing posterior probabilities over θ.

Maximum Likelihood Estimator (MLE)

- Consider the following parametric problem

$$X \sim \pi_\theta(x) = \pi(x \mid \theta), \theta \in \mathbb{R}^k$$

- Assume that the observations x_j are obtained independently, i.e. that $X_1, X_2, ..., X_N$, are i.i.d. and x_j is a realization of X_j

- Independency:

$$\pi(x_1, x_2, ..., x_N \mid \theta) = \pi(x_1 \mid \theta)\pi(x_2 \mid \theta)...\pi(x_N \mid \theta)$$

or, briefly

$$\pi(\mathcal{D} \mid \theta) = \prod_{j=1}^{N} \pi(x_j \mid \theta)$$

where

$$\mathcal{D} = \{x_1, x_2, ..., x_N\}$$

Maximum Likelihood Estimator (MLE)

- Maximum likelihood estimator (MLE) of θ = parameter value that maximizes the probability of the outcome:

$$\theta_{ML} = \arg\max_{\theta} \prod_{j=1}^{N} \pi(x_j \mid \theta)$$

- Define the negative of the log-likelihood as

$$L(\mathcal{D} \mid \theta) = -\log\left[\pi(\mathcal{D} \mid \theta)\right]$$

- Minimizer of $L(\mathcal{D} \mid \theta)$ = maximizer of $\pi(\mathcal{D} \mid \theta)$

MLE – Gaussian Model

- For a Gaussian model,

$$\pi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right), \quad \theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

the Likelihood function is given as follows:

$$\exp\left(-L(\mathcal{D} \mid \theta)\right) = \prod_{j=1}^{N} \pi(x_j \mid \theta) = \frac{1}{\left(2\pi\theta_2\right)^{N/2}} \exp\left(-\frac{1}{2\theta_2} \sum_{j=1}^{N} (x_j - \theta_1)^2 \right) =$$

$$= \exp\left(-\frac{1}{2\theta_2} \sum_{j=1}^{N} (x_j - \theta_1)^2 - \frac{N}{2} \log(2\pi\theta_2) \right) \cdots \Rightarrow$$

$$L(\mathcal{D} \mid \theta) = \frac{1}{2\theta_2} \sum_{j=1}^{N} (x_j - \theta_1)^2 + \frac{N}{2} \log(2\pi\theta_2)$$

MLE – Gaussian Model

- The gradient of the Likelihood function is then:

$$\nabla_\theta L(\mathcal{D}\,|\,\theta) = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} \\ \frac{\partial L}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\theta_2}\sum_{j=1}^{N} x_j + \frac{N}{\theta_2}\theta_1 \\ -\frac{1}{2\theta_2^2}\sum_{j=1}^{N}(x_j-\theta_1)^2 + \frac{N}{2\theta_2} \end{bmatrix} = 0$$

- This gives:

$$\mu^{mle} = \theta_{ML,1} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

$$\sigma^2_{mle} = \theta_{ML,2} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \theta_{ML,1})^2$$

- These estimates agree with what we predicted in an earlier lecture with the law of large numbers.

MLE for the Gaussian Distribution

- So for the Gaussian distribution,

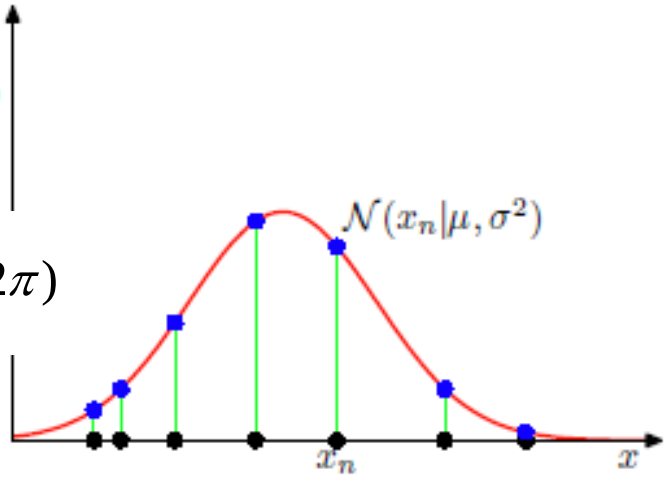
$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

the likelihood function is

$$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$$

and the log-likelihood takes the form*

$$\ln p(\mathcal{D} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$



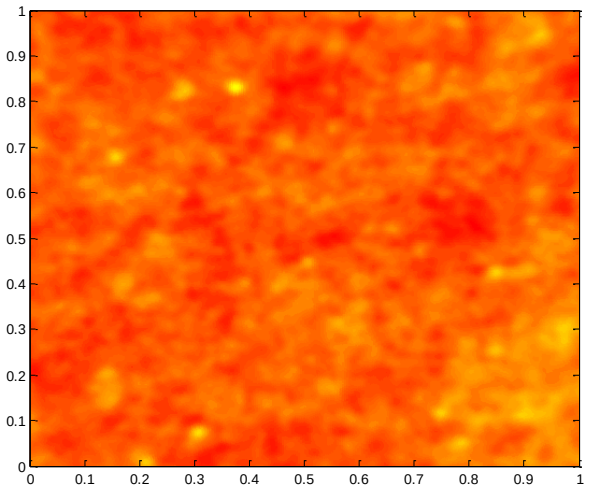
- The maximum likelihood solution is

$$\mu_{ML} = \frac{\sum_{n=1}^{N} x_n}{N}, \quad \sigma^2_{ML} = \frac{\sum_{n=1}^{N}(x_n-\mu_{ML})^2}{N}$$

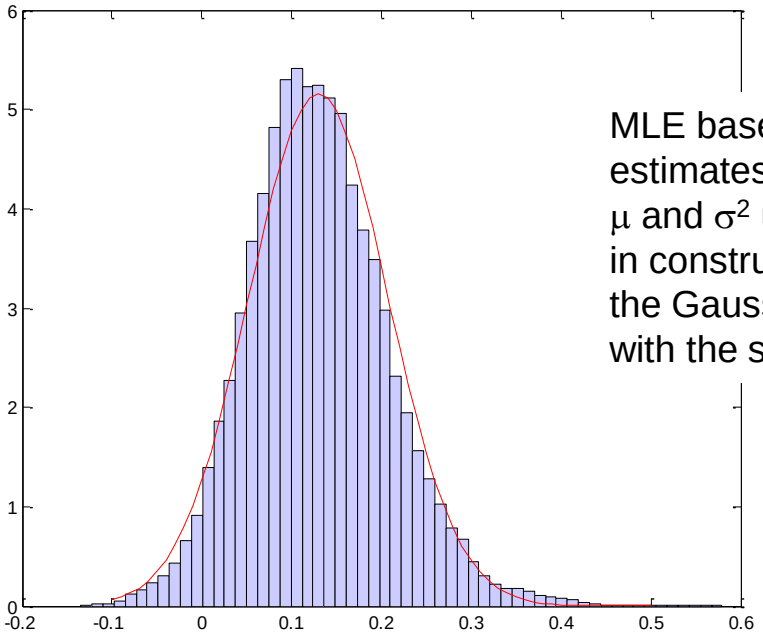
* We work often with log-likelihood to avoid underflow (taking products of small probabilities) and for simplifying the algebra.

Datasets: CMBData

- CMBdata: Spectral representation of the cosmological microwave background (CMB), i.e. electromagnetic radiation from photons back to 300,000 years after the Big Bang, expressed as difference in apparent temperature from the mean temperature.[1]



CMBdata



Matlab implementation

Normal estimation

From Bayesian Core , J.M. Marin and C.P. Roberts, Chapter 2 (available on line)

Unbiased Estimators

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameter.

- *If* $x_1, x_2, \ldots x_N \sim (i.i.d)\ \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbb{E}[\mu^{mle}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \mu$$

Thus μ^{mle} is an unbiased estimator

Biased Estimators

- An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameter.

- *If* $x_1, x_2, \ldots x_N \sim$(i.i.d) $\mathcal{N}(\mu,\sigma^2)$ then

$$\mathbb{E}\left[\sigma^2_{mle}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu^{mle})^2\right] =$$

$$\mathbb{E}\left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{j=1}^{N} x_j\right)^2\right] = \left(1-\frac{1}{N}\right)\sigma^2 \neq \sigma^2$$

Thus σ^2_{mle} is a biased estimator

MLE for a Gaussian Distribution

$$\mu_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\text{Sample mean}}, \quad \sigma^2_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2}_{\text{Sample variance wrt ML mean (not the exact mean)}}$$

- The maximum likelihood solutions μ_{ML}, σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Consider the expectations of these quantities with respect to the data set values, which come from a Gaussian.
- Using the point estimates above we showed that :

$$\mathbb{E}\left[\mu_{ML}\right] = \mu, \quad \mathbb{E}\left[\sigma^2_{ML}\right] = \frac{N-1}{N}\sigma^2$$

In this derivation you need to use :

$$\mathbb{E}\left[x_i x_j\right] = \mathbb{E}\left[x_i\right]\mathbb{E}\left[x_j\right] = \sigma^2 \quad \text{for } i \neq j$$

$$\mathbb{E}\left[x_i^2\right] = \sigma^2 + \mu^2$$

- The MLE approach thus underestimates the variance (bias) – this is in the root of the over-fitting problem.

Unbiased Estimate of Variance

- *If* $x_1, x_2, \ldots x_N \sim (i.i.d)\ \mathcal{N}(\mu, \sigma^2)$ then

$$\mathbb{E}\left[\sigma^2_{mle}\right] = \mathbb{E}\left[\frac{1}{N}\left(\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{j=1}^{N} x_j\right)^2\right] = \left(1 - \frac{1}{N}\right)\sigma^2 \neq \sigma^2$$

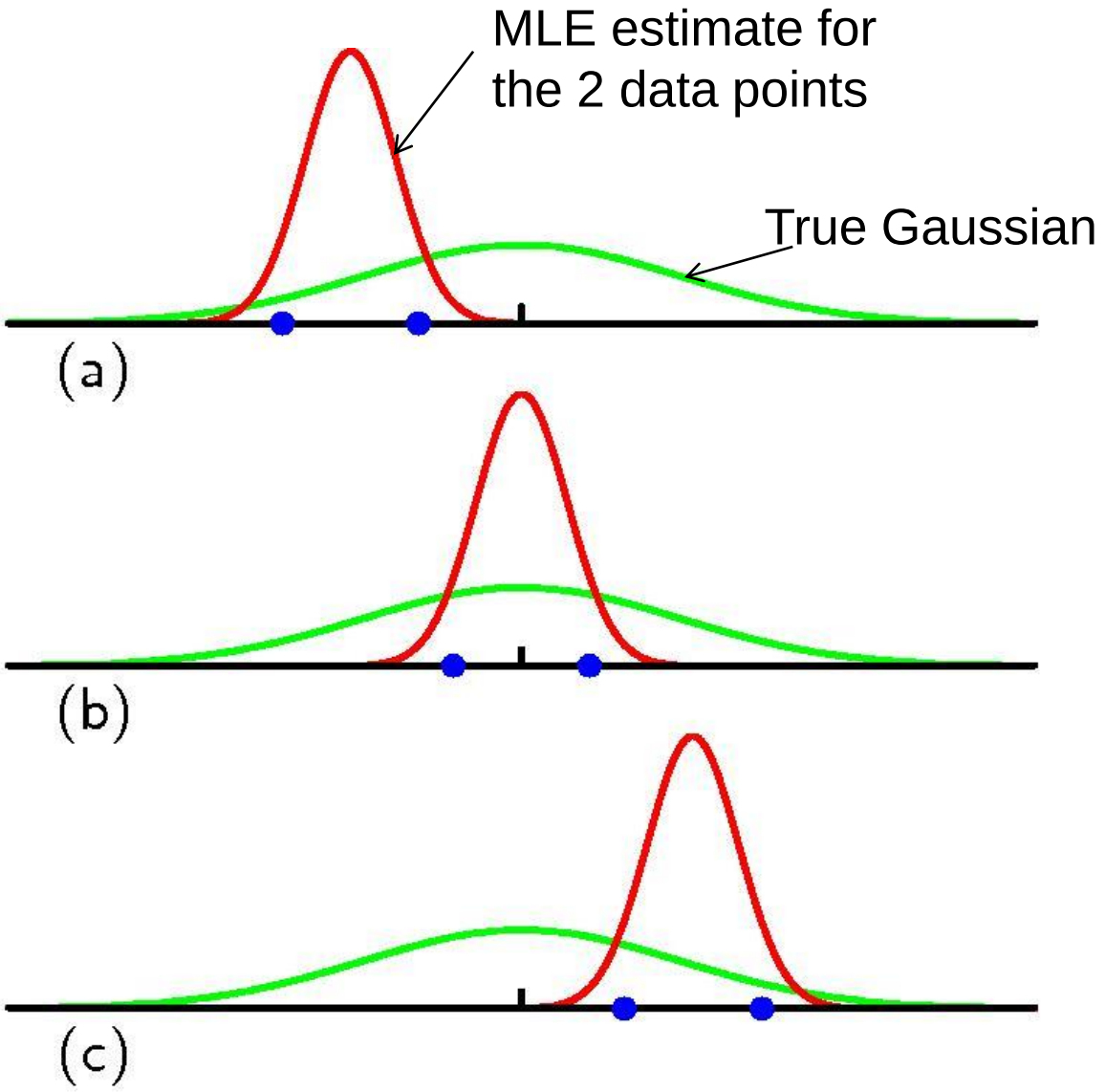
- So define

$$\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{mle}}{\left(1 - \frac{1}{N}\right)} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu^{mle})^2 \Rightarrow \mathbb{E}\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$$

- The two estimates are nearly the same for large *N*

Bias in MLE

- In the schematic from Bishop's PRML, we consider 3 cases each with 2 data points extracted from the true Gaussian.
- The mean of the three distributions predicted via MLE (i.e. averaged over the data) is correct.
- However, the variance is underestimated since it is a variance with respect to the sample mean and NOT the true mean.



Poisson Distribution

- Recall the Poisson (discrete) distribution for

$$N \in \{0, 1, 2, ..., \infty\}$$

$$P(N = n) = \pi_{Poisson}(n \mid \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

- The mean and the variance are both equal to θ.

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} n \pi_{Poisson}(n \mid \theta) = \theta,$$

$$\mathbb{E}\left[\left(N - \theta \right)^2 \right] = \theta$$

MLE – Poisson Model

- Consider the following parametric model:

$$\pi(n \mid \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

- We sample independently $\mathcal{D} = \{n_1, n_2, .., n_N\}, n_k \in \mathbb{N}$. The likelihood is:

$$\pi(\mathcal{D} \mid \theta) = \prod_{k=1}^{N} \pi(n_k) = e^{-N\theta} \prod_{k=1}^{N} \frac{\theta^{n_K}}{n_k!}$$

- The negative log likelihood function is then:

$$L(\mathcal{D} \mid \theta) == -\log \pi(\mathcal{D} \mid \theta) = \sum_{k=1}^{N} \left(\theta - n_k \log \theta + \log n_k!\right)$$

- Taking the derivative with respect to θ and setting it to zero:

$$\frac{\partial}{\partial \theta} L(\mathcal{D} \mid \theta) = \sum_{k=1}^{N} \left(1 - \frac{n_k}{\theta}\right) = 0 \Rightarrow \qquad \theta_{ML} = \frac{1}{N} \sum_{k=1}^{N} n_k$$

- However note that the Law of Large Numbers predicts:

$$\text{var}(N) \approx \frac{1}{N} \sum_{k=1}^{N} \left(n_k - \frac{1}{N} \sum_{j=1}^{N} n_j \right)^2 \neq \theta_{ML}$$

MLE – Poisson Model

- Assume that θ is known a prior to be large. In this case, we can use the Gaussian approximation of Poisson distribution (result derived in an earlier lecture):

$$\prod_{j=1}^{N} \pi_{Poisson}(n_j \mid \theta) \approx \left(\frac{1}{2\pi\theta}\right)^{N/2} \exp\left(-\frac{1}{2\theta}\sum_{j=1}^{N}\left(n_j-\theta\right)^2\right) =$$

$$= \left(\frac{1}{2\pi}\right)^{N/2} \exp\left(-\frac{1}{2}\left[\frac{1}{\theta}\sum_{j=1}^{N}\left(n_j-\theta\right)^2 + N\log\theta\right]\right)$$

$$L(\mathcal{D} \mid \theta) = \frac{1}{\theta}\sum_{j=1}^{N}\left(n_j-\theta\right)^2 + N\log\theta$$

$$\theta_{ML}: \quad \frac{\partial}{\partial\theta} L(\mathcal{D} \mid \theta) = -\frac{1}{\theta^2}\sum_{j=1}^{N}\left(n_j-\theta\right)^2 - \frac{2}{\theta}\sum_{j=1}^{N}\left(n_j-\theta\right) + \frac{N}{\theta} = 0 \Rightarrow$$

- An approximation for

$$\theta = \left(\frac{1}{4} + \frac{1}{N}\sum_{j=1}^{N} n_j^2\right)^{1/2} - \frac{1}{2} \neq \frac{1}{N}\sum_{j=1}^{N} n_j \; (\textit{result from the exact density})$$

MLE for the Multinomial Distribution

- Suppose categorical arity-*n* inputs $x_1, x_2, \ldots x_N$~(i.i.d.) from a multinomial

$$\mathcal{M}(p_1, p_2, \ldots p_N)$$

where

$$P(x_k=j|\mathbf{p})=p_j$$

- What is the MLE of $\mathbf{p}=(p_1, p_2, \ldots p_N)$?

$$f(x_1, x_2, \ldots, x_N \mid p_1, p_2, \ldots, p_N) = \frac{N!}{\prod x_i!} \prod p_i^{x_i} = \binom{N}{x_1, x_2, \ldots, x_N} p_1^{x_1} p_2^{x_2} \ldots p_N^{x_N}$$

- The penalized log-likelihood is given as:

$$\ell(p_1, p_2, \ldots, p_N) = \log N! - \sum_{i=1}^{N} \log x_i! + \sum_{i=1}^{N} x_i \log p_i + \underbrace{\lambda\left(1 - \sum_{i=1}^{N} p_i\right)}_{\text{Lagrange multiplier enforced constraint}}$$

- Differentiation gives the expected result:

$$p_i^{mle} = \frac{x_i}{N}$$

MLE and Weighted Least Squares

- Consider a multivariate Gaussian model,

$$X \sim \mathcal{N}(x_0, \Gamma)$$

where $x_0 \in \mathbb{R}^n$ is unknown, $\Gamma \in \mathbb{R}^{n \times n}$ is a known symmetric positive definite matrix

- Assume that x_0 depends on hidden parameters $z \in \mathbb{R}^k$ through a linear equation (Model Reduction Approach, *k*<<*n*)

$$x_0 = Az,\ A \in \mathbb{R}^{n \times k},\ z \in \mathbb{R}^k$$

- Every time you introduce model reduction, you also introduce model errors.

- **We consider the inverse problem of computing *z from realizations of X.***

MLE and Weighted Least Squares

- Our problem can also be written by considering noisy observations as:

$$X = Az + E, \quad E \sim \mathcal{N}(0, \Gamma)$$

- Note that:

$$\mathbb{E}[X] = Az + \mathbb{E}[E] = Az = x_0$$

$$\text{cov}[X] = \mathbb{E}\left[XX^T\right] = \mathbb{E}\left[(Az+E)(Az+E)^T\right] = \mathbb{E}\left[EE^T\right] = \Gamma$$

- The probability density of X given z, is:

$$\pi(x \mid z) = \frac{1}{(2\pi)^{n/2}\det(\Gamma)^{1/2}} \exp\left(-\frac{1}{2}(x-Az)^T\Gamma^{-1}(x-Az)\right)$$

- Assume independent observations $\mathcal{D} = \{x_1, x_2, .., x_N\}, x_j \in \mathbb{R}^n$

- The likelihood function is then $\prod_{j=1}^{N} \pi(x_j \mid z) \sim \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}(x_j - Az)^T\Gamma^{-1}(x_j - Az)\right\}$

- Now minimize $L(\mathcal{D}|z)$:

$$L(\mathcal{D} \mid z) = \frac{1}{2}\sum_{j=1}^{N}(x_j - Az)^T\Gamma^{-1}(x_j - Az) = \frac{N}{2} z^T\left[A^T\Gamma^{-1}A\right]z - z^T\left[A^T\Gamma^{-1}\sum_{j=1}^{N}x_j\right] + \frac{1}{2}\sum_{j=1}^{N}x_j^T\Gamma^{-1}x_j$$

MLE and Weighted Least Squares

$$L(\mathscr{D}\,|\,z)=\frac{1}{2}\sum_{j=1}^{N}(x_j-Az)^T\Gamma^{-1}(x_j-Az)=\frac{N}{2}z^T\left[A^T\Gamma^{-1}A\right]z-z^T\left[A^T\Gamma^{-1}\sum_{j=1}^{N}x_j\right]+\frac{1}{2}\sum_{j=1}^{N}x_j^T\Gamma^{-1}x_j$$

- Taking the gradient wrt z equal to zero:

$$\nabla_z L(\mathscr{D}\,|\,z)=N\left[A^T\Gamma^{-1}A\right]z-\left[A^T\Gamma^{-1}\sum_{j=1}^{N}x_j\right]=0\Rightarrow$$

$$\left[A^T\Gamma^{-1}A\right]z=A^T\Gamma^{-1}\bar{x},\quad where\quad \bar{x}=\frac{1}{N}\sum_{j=1}^{N}x_j$$

- The existence of the solution of this system depends on the matrix $A\in\mathbb{R}^{n\times k}$
- For the case of one observation, i.e. $\mathscr{D}=\{x\}$: $L(x\,|\,z)=(x-Az)^T\Gamma^{-1}(x-Az)$
- Using $\Gamma=UDU^T,\Gamma^{-1}=(D^{-1/2}U^T)^T(D^{-1/2}U^T)=W^TW,\quad W=D^{-1/2}U^T$, we can finally write:

$$L(x\,|\,z)=\left\|W(Az-x)\right\|^2$$

- The MLE minimization problem is a weighted least squares problem!

MLE for Multivariate Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N \sim$(i.i.d) $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- We do not know $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$
- MLE problem: For which $\theta = (\mu,\Sigma)$ is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k \; or \; in \; component \; form \; \mu_i^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_{ki}, 1 \le i \le m$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)^T \; or \; in \; component \; form$$

$$\sigma_{ij}^{mle} = \frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$

$$\boldsymbol{\Sigma}^{\text{unbiased}} = \frac{\boldsymbol{\Sigma}^{mle}}{1-\frac{1}{N}} = \frac{1}{N-1}\sum_{k=1}^{N}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Note that $\boldsymbol{\Sigma}^{mle}$ comes to be symmetric non-negative definite

Frequentist Vs Bayesian Approach to Statistical Inference

- Frequentist approach: In all previous examples, the data was considered as sample from a parametric probability density and the underlying parameters that we seek to compute from the data were deterministic.

- Bayesian approach: Any quantity that is not known is taken as random variable - randomness implies lack of information.

Inverse Problems

- Our inverse problem of interest is to estimate a parameter $x \in \mathbb{R}^n$ that we cannot observe directly.
- Some information may be known about $x, e.g.\ x \in B.$
- We observe another vector $y \in \mathbb{R}^k$ that depends on x through a mathematical model:

$$y = f(x)$$

- Find an estimate x such that

minimize $\|y - f(x)\|$ subject to constraint $x \in B$

A Bayesian Approach to Inverse Problems

- In general, we have some a priori knowledge about the unknown
- We also have a mathematical model (forward solver) that explains the observations, with all uncertainties included
- In a Bayesian inference approach,
  - we express *x* as a parameter that defines the distribution of y (likelihood model)
  - Incorporate prior information into the model (prior model).

A Bayesian Approach to Inverse Problems

- In a Bayesian approach, everything that is not known is taken as a random variable.

- A typical approach is to consider one unknown at a time using conditioning, e.g.

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

- If a variable is of no interest to the analysis, integrate it out using marginalization:

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

Likelihood

- Assuming that we know the unknown *x*, how would the measurement be distributed?

- Randomness of the measurement *y*, provided that *x* is known, is due to

  - measurement noise

  - incompleteness in the computational model:

    - ✓ discretization errors

    - ✓ approximation of reality through a model

    - ✓ various (unknown) nuisance parameters

Additive Noise Model

- Assume a functional dependence, $y = f(x)$ with errors in the observations. Here $x \in \mathbb{R}^n$ is the unknown and $y \in \mathbb{R}^m$ the observable.

- A frequently used model is the additive noise model,

$$Y = f(X) + E$$

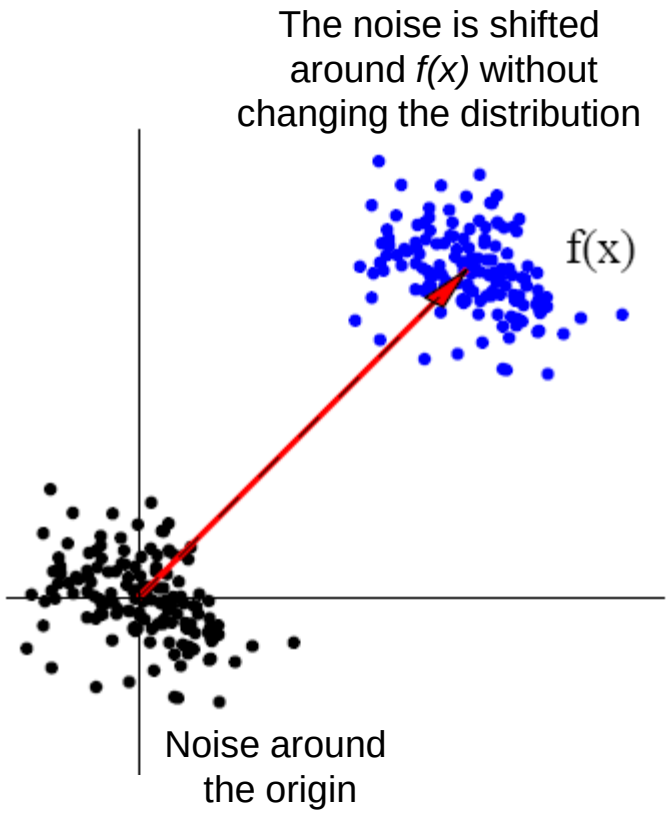
where the distribution of the error is

$$E \sim \pi_{noise}(e)$$

- We assume π_{noise} to be known.

- If E and X are mutually independent,

$$\pi(y|x) = \pi_{noise}(y - f(x))$$



Additive Noise Model

- The noise distribution may depend on unknown parameters θ:

$$\pi_{noise}(e) = \pi_{noise}(e|\theta)$$

- The likelihood in this case is given as:

$$\pi(y|x,\theta) = \pi_{noise}(y-f(x)|\theta)$$

- If E is zero mean Gaussian with unknown variance σ^2, i.e.

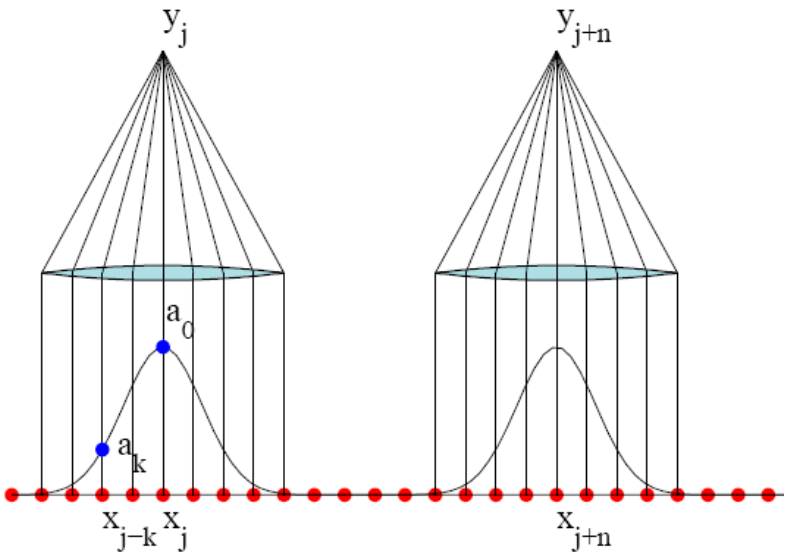
$$E \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{I} \in \mathbb{R}^{m \times m}$ is the identity matrix, the likelihood is

$$\pi(y \mid x, \sigma^2) = \frac{1}{(2\pi)^{m/2}\,\sigma^m} \exp\left(-\frac{1}{2\sigma^2} \|y - f(x)\|^2 \right)$$

An Example of Likelihood Calculation

- Consider a device consisting of a collecting lens and a photon counter, where the photons come from *N* emitting sources.
- Let the average photon emission/observation time = $x_j, 1 \le j \le N$
- The average total count (defined by the geometry, *L*, of the lens) is taken as the weighted sum (weights a_j) of the individual contributions.



$$\bar{y}_j = \mathbb{E}[Y_j] = \sum_{k=L}^{k=-L} a_k x_{j-k}$$

We assume:
$x_j = 0$ if $j < 1$ or $j > N$

D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007

Likelihood Calculation

- Accounting for all source points, we can write the following:

$$\bar{y} = \mathbb{E}\left[Y\right] = Ax$$

- $A \in \mathbb{R}^{n \times n}$ is a Toeplitz matrix with L defining the bandwidth of the matrix.

$$A = \begin{bmatrix} a_0 & a_{-1} & \dots & a_{-L} & & \\ a_1 & a_0 & & & \ddots & \\ \ddots & & \ddots & & & a_{-L} \\ a_L & & & \ddots & & \ddots \\ & & & & a_0 & a_{-1} \\ & & a_L & \ddots & a_1 & a_0 \end{bmatrix}$$

Likelihood Calculation

- The photon counting process is a Poisson process:

$$Y_j \sim \mathcal{Poisson}\Big(\big(Ax\big)_j\Big)$$

- Explicitly we can write:

$$\pi(y_j \mid x) = \frac{\big(Ax\big)_j^{y_j}}{y_j!}\exp\Big(-\big(Ax\big)_j\Big)$$

- We assume that the consecutive measurements are independent, thus $Y \in \mathbb{R}^n$ has the density

$$\pi(y \mid x) = \prod_{j=1}^{N} \pi(y_j \mid x) = \prod_{j=1}^{N} \frac{\big(Ax\big)_j^{y_j}}{y_j!}\exp\Big(-\big(Ax\big)_j\Big)$$

$$Y \sim \mathcal{Poisson}(Ax)$$

Approximate the Poisson Process With a Gaussian

- Approximate the Poisson process with a Gaussian process for a high count of photons (validity of approximation shown in an earlier lecture)

$$\pi(y \mid x) = \prod_{l=1}^{N} \left(\frac{1}{2\pi (Ax)_l} \right)^{1/2} \exp\left(-\frac{1}{2(Ax)_l} \left(y - (Ax)_l \right)^2 \right) =$$

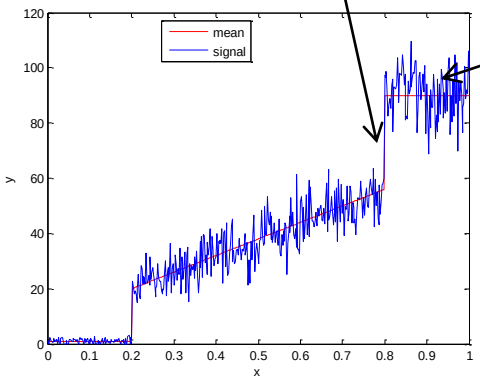
$$= \left(\frac{1}{(2\pi)^N \det \Gamma} \right)^{1/2} \exp\left(-\frac{1}{2} (y - Ax)^T \Gamma^{-1} (y - Ax) \right)$$

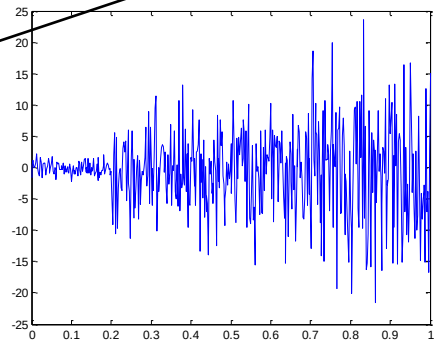
where: $\Gamma = \Gamma(x) = diag(Ax)$

- Note that the higher the signal, the higher the noise (recall for Poisson process, the mean and the variance are the same)

Average piece-wise linear signal $\bar{x}$

Realization of a Poisson process with $\bar{x}$ as its mean





Click here for a MatLab implementation

Difference between the noisy signal and $\bar{x}$

MLE Not a Useful Estimator for Convolution Problems

- Computing the MLE estimate of *x* by differentiating wrt to *x* the likelihood below is difficult (Γ depends on x, $\Gamma = \Gamma(x) = diag(Ax)$, and there is high sensitivity to noise in the data).

$$\pi(y \mid x) = \left(\frac{1}{(2\pi)^N \det \Gamma} \right)^{1/2} \exp\left(-\frac{1}{2} (y - Ax)^T \Gamma^{-1} (y - Ax) \right), \Gamma = diag(Ax)$$

- This is a typical issue with convolution problems.

- One needs either a fully Bayesian treatment (prior on x) or using classical regularization techniques.

Change of Variables

- In calculating the likelihood function, we often have to deal with computing distributions when changing variables.

- Consider two random variables *X and Y in* $\mathbb{R}^n$

$$Y = f(X)$$

where *f* is a differentiable function. Assume that the probability distribution of *Y* is known:

$$\pi(y) = p(y)$$

- What is the probability density of *X?*

$$\pi(y)dy = p(y)dy = p(f(x))|det(Df(x))|dx$$

from which we can conclude that

$$\pi(x) = p(f(x))|det(Df(x))|$$

Change of Variables

- In calculating the likelihood function, we often have to deal with computing distributions when changing variables.

- Consider two random variables *X and Y in* $\mathbb{R}^n$

$$Y = f(X)$$

where *f* is a differentiable function. Assume now that the probability distribution of *X* is known:

$$\pi(x) = p(x)$$

- What is the probability density of *Y?*

- *Using* $X = f^{-1}(y)$ *and the result from the earlier slide:*

$$\pi(y) = p\left(f^{-1}(y)\right)\left|\det \frac{df^{-1}(y)}{dy}\right| = \frac{p\left(f^{-1}(y)\right)}{\left|\det \frac{df(x)}{dx}\right|}$$

Change of Variables: Example

- Consider an amplifier input signal *f(t)* amplified by a factor $\alpha > 1$ that fluctuates.

- We model the output signal as follows:

$$g(t) = \alpha f(t), \ 0 \le t \le T$$

or in discrete form:

$$x_j = f(t_j), y_j = g(t_j), 0 < t_1 < t_2 ... < t_n = T$$

- We denote the amplification at $t = t_j$ as a_j *:*

$$y_j = \alpha_j x_j, 1 \le j \le n$$

Multiplicative Noise

- The stochastic extension of the input/output signal relation is:

$$Y_j = A_j X_j, 1 \le j \le n$$

or in the vector notation as (component wise relation)

$$Y = A.X$$

- Let us assume that A has the probability density (multiplicative noise)

$$A \sim \pi_{noise}(a)$$

- Then the likelihood density for Y, conditioned on $X = x$, $\pi(y \mid x)$ is

$$\pi_{noise}(a)\, da = \underbrace{\pi_{noise}(a(y))\,|\det Da(y)|}_{\pi(y|x)} dy$$

$$\pi(y \mid x) = \frac{1}{x_1 x_2 \ldots x_n} \pi_{noise}\left(\frac{y_\bullet}{x}\right) \quad (normalized)$$

component wise division

Multiplicative Noise

- We can also derive this likelihood as follows:

- From $y = a.x$, and $a = \frac{y_\bullet}{x}\,(x\ fixed)$ or $a_j = \frac{y_j}{x_j} \Rightarrow da_j = \frac{dy_j}{x_j}$

- We can now write:

$$p(a)da = p(a)da_1 \ldots da_n = p(\frac{y_\bullet}{x})\frac{dy_1}{x_1}\ldots\frac{dy_n}{x_n} = \underbrace{\left(\frac{1}{x_1 \ldots x_n}\, p(\frac{y_\bullet}{x})\right)}_{\pi(y|x)} dy_1 \ldots dy_n$$

$$\pi(y \mid x) = \frac{1}{x_1 \ldots x_n}\, p(\frac{y_1}{x}) \ldots p(\frac{y_n}{x})$$

Change of Variables: Example

- Let us now consider again the multiplicative noise case but now with all the variables being positive, and *A* is log-normally distributed:

$$W_i = \log A_i \sim \mathcal{N}(w_0, \sigma^2), \quad w_0 = \log \alpha_0$$

The A_i components are mutually independent.

- The probability distributions transform as densities, not as functions.

$$P\{W_i = \log A_i < t\} = P(A_i < e^t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} e^{-\frac{1}{2\sigma^2}(w_i - w_0)^2} dw_i$$

Change of Variables: Example

- Change of variables:

$$w_i = \log a_i \Rightarrow dw_i = \frac{1}{a_i} da_i$$

- Substitute $w_0 = \log \alpha_0$

$$P\{W_i < t\} = P(A_i < e^t) = \int_{-\infty}^{t} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{a_i} \exp\left(-\frac{1}{2\sigma^2} \left(\log \frac{a_i}{\alpha_0} \right)^2 \right)}_{\substack{\pi(a_i) \\ 1D\ \log-normal \\ distribution}} da_i$$

- Consider independent components and use

$$\pi(y_i \mid x_i) = \frac{da_i}{dy_i} \pi_{noise}\left(\frac{y_i}{x_i} \right) = \frac{1}{x_i} \pi_{noise}\left(a_i \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x_i a_i} \exp\left(-\frac{1}{2\sigma^2} \left(\log \frac{a_i}{\alpha_0} \right)^2 \right)$$

to derive:

$$\pi(y \mid x) = \pi(y_1 \mid x_1)\pi(y_2 \mid x_2)...\pi(y_n \mid x_n) = \prod_{y_i}$$

$$\left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \frac{1}{y_1 ... y_n} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(\log \frac{y_j}{\alpha_0 x_j} \right)^2 \right)$$

Likelihood Calculation: Additive Poisson & Gaussian Noise

- Let us consider the same example with additive Poisson and Gaussian noise:

$$Y = Z + E, \quad Z \sim \mathcal{Poisson}(Ax), \quad E \sim \mathcal{N}(0, \sigma^2 I)$$

- First step: assume that $X = x$ and $Z = z$ are known, giving

$$\pi(y_j \mid z_j, x) \sim \exp\left(-\frac{1}{2\sigma^2} \left(y_j - z_j \right)^2 \right)$$

- Conditioning: $\pi(y_j, z_j \mid x) = \pi(y_j \mid z_j, x)\pi(z_j \mid x)$

- The value of z_j (integer) is not of interest here, so the needed likelihood (function of x) is:

$$\pi(y_j \mid x) = \sum_{z_j=0}^{\infty} \pi(y_j, z_j \mid x) = \sum_{z_j=0}^{\infty} \pi(z_j \mid x) \exp\left(-\frac{1}{2\sigma^2} \left(y_j - z_j \right)^2 \right)$$

- One can further simplify by using the Gaussian approximation to $\pi(z_j \mid x)$

Back to Bayes' Rule

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\pi(x)}$$

- $\pi(\theta)$: prior distribution
  - contains the prior information about the unknown parameters of hypothesis θ. The prior distribution can be subjective or objective.

- $f(x|\ \theta)$: likelihood function
  - It provides the link between the hypothesis one wants to examine and the actual evidence one has

- $\pi(\theta|x)$: posterior distribution
  - In contrast to classical statistical inference techniques, the result of Bayesian inference is a probability distribution

Posterior Densities

Fundamental identity:

$$\pi(\theta, x) = \pi_{prior}(\theta)\, f(x \mid \theta) = \pi(x)\pi(\theta \mid x)$$

Bayes' formula

$$\pi(\theta \mid x) = \frac{\pi_{prior}(\theta)\, f(x \mid \theta)}{\pi(x)},\ x = observed$$

Here $\pi(\theta \mid x)$ is the posterior density

The posterior density is the Bayesian solution of the inverse problem.

Posterior Example: Linear Inverse Problem

- Let us revisit the example discussed earlier with *z* now becoming a random variable. Consider a linear inverse problem, additive noise (*z* is the main unknown)

$$x = Az + e, \quad z \in \mathbb{R}^k, \quad x, e \in \mathbb{R}^n, \quad A \in \mathbb{R}^{n \times k}$$

- Stochastic extension $X = AZ + E$

- Assume that Z *and E are independent and Gaussian,*

$$Z \sim \mathcal{N}(0, \gamma^2 \Gamma), \quad E \sim \mathcal{N}(0, \sigma^2 I)$$

- *The prior density is*

$$\pi_{prior}(z \mid \gamma) \propto \frac{1}{\gamma^n} \exp\left(-\frac{1}{2\gamma^2} z^T \Gamma^{-1} z \right)$$

- *Observe that:* $\det(\gamma^2 \Gamma) = \gamma^{2n} \det(\Gamma)$

- Also the likelihood is: $\pi(x \mid z) \propto \exp\left(-\frac{1}{2\sigma^2} \|x - Az\|^2 \right)$

Posterior: MAP Estimate

- Using Bayes' formula, the posterior is finally given as:

$$\pi(z \mid x,\gamma) \sim \pi_{prior}(z \mid \gamma)\, f(x \mid z) \sim \frac{1}{\gamma^n} \exp\left(\underbrace{-\frac{1}{2\gamma^2} z^T \Gamma^{-1} z - \frac{1}{2\sigma^2} \left\| x - Az \right\|^2}_{-V(z|x,\gamma)} \right)$$

$$\pi(z \mid x,\gamma) \sim \frac{1}{\gamma^n} \exp\left(-\underbrace{V(z \mid x,\gamma)}_{\substack{Tichonov \\ functional}} \right)$$

- *At this point, we are only interested in point estimates of z, in particular the MAP estimate: the value of z that maximizes the posterior.*

The Tikhonov Functional

- The matrix Γ is symmetric and positive definite. Using Cholesky factorization:

$$\Gamma^{-1} = R^T R$$

  Here *R* is an upper triangular matrix.

- Then: $z^T \Gamma^{-1} z = z^T R^T R z = \|Rz\|^2$

- From which it follows the following Tichonov functional *T*:

$$T(z) \equiv 2\sigma^2 V(z \mid x, \gamma) = \|x - Az\|^2 + \delta^2 \|Rz\|^2, \quad \delta = \frac{\sigma}{\gamma}$$

- The MAP estimate starts now looking identical to least squares optimization with Tichonov regularization – the only difference is now that the regularization parameter is defined from the modeling errors (likelihood) and the prior model on *z!*

Maximum a Posterior Estimator (MAP)

- The Bayesian analogue of the maximum likelihood estimator

$$z_{MAP} = \arg\max_{z} \pi(z \mid x)$$

- Or equivalently

$$z_{MAP} = \arg\min_{z} V(z \mid x), \quad V(z \mid x) = -\log \pi(z \mid x)$$

- Here:

$$z_{MAP} = \arg\min_{z} \left(\|x - Az\|^2 + \delta^2 \|Rz\|^2 \right)$$

Maximum a Posterior Estimator (MAP)

- The Maximum a Posterior estimator is the penalized least squares solution of the problem

$$Az = x$$

- An equivalent characterization of the MAP estimator is given as the minimization of:

$$\|x - Az\|^2 + \delta^2 \|Rz\|^2 = \left\| \begin{bmatrix} x \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \delta R \end{bmatrix} z \right\|^2$$

- Thus the MAP estimate is the least squares solution of

$$\begin{bmatrix} A \\ \delta R \end{bmatrix} z = \begin{bmatrix} x \\ 0 \end{bmatrix}$$

- But note that a Bayesian approach to inverse problems is much more than point (MAP) estimates. We have available the distribution $p(z|x)$ to expore!

Empirical Priors

- ✓ Assume that we try to determine the parameter *z* (using indirect measurements *x)* - as in our earlier model reduction problem.
- ✓ Assume that we have some previous measurements of the unknown parameter z directly,

$$\mathcal{Z} = \{z_1, z_2, ..., z_N\}$$

- ✓ Think of these as realizations of a random variable with an unknown distribution.
- ❑ Non-parametric approach: Look at a histogram based on $\mathcal{Z}$.
- ❑ Parametric approach: Justify a parametric model, find the maximum likelihood estimate of the model parameters.

D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007

ML Estimate: Empirical Prior

- Let us assume that $Z \sim \mathcal{N}(z_0, \sigma^2)$

- From previous analysis, the ML estimate for z_0 *is*

$$z_{0,ML} = \frac{1}{N}\sum_{j=1}^{N} z_j$$

- *And for σ^2:*

$$\sigma^2_{ML} = \frac{1}{N}\sum_{j=1}^{N}\left(z_j - z_{0,ML}\right)^2$$

Empirical Bayes Approach to Prior

- We postulate that the unknown *parameter Z* is a random variable, whose probability distribution is denoted as $\pi_{pr}(Z)$ and called the prior distribution

- *We* use the parametric model

$$\pi_{pr}(Z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z - z_0)^2 \right)$$

where z_0 *and* σ^2 are determined experimentally from the data $\mathcal{Z}$ by the formulas in the earlier slide.

This approach where the prior is defined through previous experience, is called ***empirical Bayes approach.***

Smoothness Priors

- We talked earlier about point estimates (MLE, MAP) and weighted regularized least squares. The following question is now posed:

- Can priors be introduced that lead to first order Tichonov regularization (one penalizing norms of derivatives) when using MAP estimates?

- We introduce next an example using a smoothness prior.

Prior Modeling: Smoothness Prior

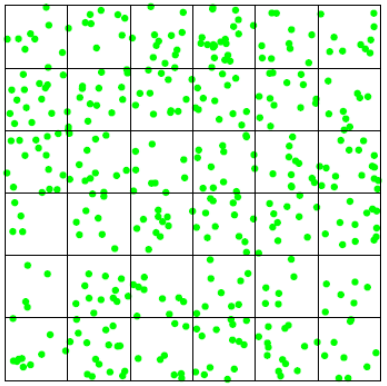
- Consider that z represents an unknown property (parameter) defined on a rectangular array of squares (e.g. with each square containing e.g. a number of bacteria there)

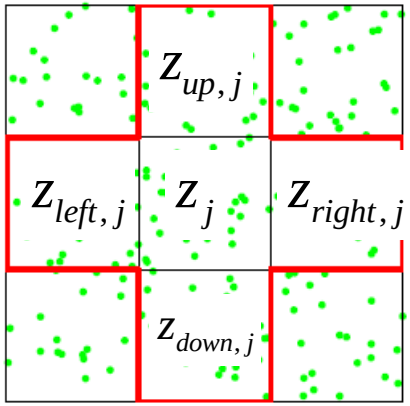
- We want to estimate the (prior) density of the bacteria from indirect measurements.

- We set up a prior model based on our belief how bacteria grow.*

- Number of bacteria in a box is taken as the average of neighbours, i.e.

$$z_j \approx \frac{1}{4}\left(z_{left,j} + z_{right,j} + z_{up,j} + z_{down,j} \right)$$

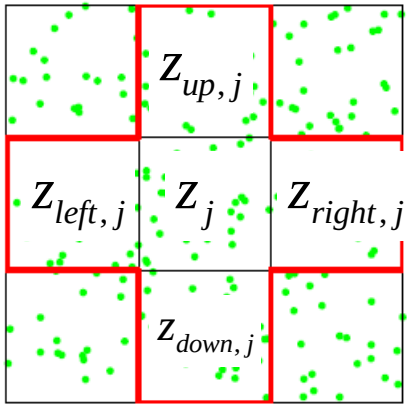




* We already have seen a similar smoothness prior used for the interpolation of noise-free data.

Prior Modeling: Smoothness Prior

- We define $z_j = 0$ *for pixels outside the square.*



- Introduce matrix $A \in \mathbb{R}^{k \times k}$, *k = number of pixels,*

(up) (down) (left) (right)

$$A(j,:) = [0 \ldots 1/4 \ldots 1/4 \ldots 1/4 \ldots 1/4 \ldots 0]$$

- Absolute certainty about our model leads to:

$$z = Az$$

- This is not correct since:

$$(I - A)z = 0 \Rightarrow z = 0, \ \text{since} \ \det(I - A) \neq 0$$

Prior Modeling: Smoothness Prior

- As a solution to this problem, we relax the model and write

$$z = Az + r;\ r = \text{uncertainty of the model}$$

- We model *r* as a random variable and postulate a distribution:

$$r \sim \pi_{\text{mod.}error}(r)$$

- From $z - Az = r$ *follows a natural prior model,*

$$\pi_{prior}(z) = \pi_{\text{mod.}error}(z - Az)$$

- The model is referred to as autoregressive Markov model, and r is an innovation process.

Prior Modeling: Smoothness Prior

- If *r* is a Gaussian variable with mutually independent and equally distributed components,

$$r \sim \mathcal{N}(0, \sigma^2 I)$$

we obtain the prior model

$$\pi_{prior}(z \mid \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|z - Az\|^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|Lz\|^2\right)$$

$$L = I - A$$

- σ^2 is usually not known – it is a parameter in the prior model for *z*. Hierarchical models to be discussed in later lectures are used in such cases.
- The form of the smoothness prior above is the one used in our previous lecture.

2nd-Order Smoothness Prior

- L is a second order derivative finite difference matrix with the structure

$$\begin{bmatrix} & -1/4 & \\ -1/4 & 1 & -1/4 \\ & -1/4 & \end{bmatrix}$$

- The model leads to what is referred to as the *second order smoothness prior.*
- Another derivation: Assume that $z_j = f(p_j)$; p_j = *point in the* j^{th} *pixel:*
- Finite difference approximation (h=discretization step) of the Laplacian of f gives:

$$-\Delta f(p_j) \approx \frac{4}{h^2}(Lx)_j$$

- This prior gives higher probability to vectors $x \in \mathbb{R}^n$ corresponding to discrete approximations of functions with a small 2nd derivative Δf .

D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007
Click here for an implementation of the smoothness prior

Gaussian Linear Models

Bayes' Theorem and Gaussian Linear Models

- Consider a linear Gaussian model: A Gaussian marginal distribution $p(\mathbf{x})$ and a Gaussian conditional distribution $p(\mathbf{y}|\mathbf{x})$ in which $p(\mathbf{y}|\mathbf{x})$ has a mean that is a linear function of $\mathbf{x}$, and a covariance which is independent of $\mathbf{x}$.

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)$$

- *We want using Bayes' rule to find $p(\mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$.*

- We start with the joint distribution over $\mathbf{z}=(\mathbf{x},\mathbf{y})$ which is quadratic in the components of $\mathbf{z}$ – *so* $p(\mathbf{z})$ is a Gaussian.

Bayes' Theorem and Gaussian Processes

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

$$\ln p(\boldsymbol{z}) = \ln p(\boldsymbol{x}) + \ln p(\boldsymbol{y} \mid \boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Lambda (\boldsymbol{x} - \boldsymbol{\mu})$$

$$-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + const$$

Covariance of the Joint Distribution

$$\ln p(\boldsymbol{z}) = -\frac{1}{2}\boldsymbol{x}^T\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\right)\boldsymbol{x} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y} + const =$$

$$= -\frac{1}{2}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix} + const$$

Only quadratic terms shown

- We can immediately write down the covariance of $\boldsymbol{z}$.

$$cov[\boldsymbol{z}] = \begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1}+\boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}$$

- In the matrix inversion we used a result from an earlier lecture.

$$\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{M}^{-1} & -\boldsymbol{M}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}^{-1} & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1}\end{pmatrix}, \quad where: \boldsymbol{M} = \boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$$

Mean of the Joint Distribution

$$\ln p(\boldsymbol{z}) = -\boldsymbol{x}^T \boldsymbol{\Lambda}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{y}^T \boldsymbol{L}\boldsymbol{b} - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{L}\boldsymbol{b} + \ldots$$

$$= \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L}\boldsymbol{b} \\ \boldsymbol{L}\boldsymbol{b} \end{pmatrix} + \ldots$$

Only linear terms shown

- We can immediately write down the mean of **z:**

$$\boldsymbol{z}^T cov[\boldsymbol{z}]^{-1} \mathbb{E}[\boldsymbol{z}] = \boldsymbol{z}^T \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L}\boldsymbol{b} \\ \boldsymbol{L}\boldsymbol{b} \end{pmatrix}$$

$$\mathbb{E}[\boldsymbol{z}] = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L}\boldsymbol{b} \\ \boldsymbol{L}\boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}$$

- It remains to find the marginal p(**y**). We can use earlier derived results.

Marginal p(y) Distribution

- Recall our earlier result for computing the marginal:

$$p\left(\boldsymbol{x}_a\right)=\mathcal{N}\left(\boldsymbol{x}_a \mid \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right)$$

- Based on our calculations:

$$\mathbb{E}[\boldsymbol{z}]=\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{A\mu}+\boldsymbol{b}\end{pmatrix} \qquad cov[\boldsymbol{z}]=\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A\Lambda}^{-1} & \boldsymbol{L}^{-1}+\boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}$$

we conclude:

$$p\left(\boldsymbol{y}\right)=\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A\mu}+\boldsymbol{b}, \boldsymbol{L}^{-1}+\boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T\right)$$

$$cov[\boldsymbol{y}]=\boldsymbol{L}^{-1}+\boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T$$

$$\mathbb{E}[\boldsymbol{y}]=\boldsymbol{A\mu}+\boldsymbol{b}$$

- Note that for $\boldsymbol{A}=\boldsymbol{I}$, $p\left(\boldsymbol{x}\right)=\mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right), p\left(\boldsymbol{y} \mid \boldsymbol{x}\right)=\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{x}+\boldsymbol{b}, \boldsymbol{L}^{-1}\right)$ the convolution of the two Gaussians gives the well known result:

$$\mathbb{E}[\boldsymbol{y}]=\boldsymbol{\mu}+\boldsymbol{b}$$

$$cov[\boldsymbol{y}]=\boldsymbol{L}^{-1}+\Lambda^{-1}$$

Conditional p(x|y) Distribution

- Recall our earlier result for computing the conditional:

$$p\left(\boldsymbol{x}_a \mid \boldsymbol{x}_b\right)=\mathcal{N}\left(\boldsymbol{x}_a \mid \boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right) \qquad \boldsymbol{\mu}_{a|b}=\boldsymbol{\mu}_a-\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\boldsymbol{x}_b-\boldsymbol{\mu}_b\right)$$

- Based on our calculations:

$$\mathbb{E}[\boldsymbol{z}]=\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{A\mu}+\boldsymbol{b}\end{pmatrix} \qquad cov[\boldsymbol{z}]=\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{LA} & \boldsymbol{L}\end{pmatrix}^{-1}=\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A\Lambda}^{-1} & \boldsymbol{L}^{-1}+\boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T\end{pmatrix}$$

we conclude: $p\left(\boldsymbol{x}\mid\boldsymbol{y}\right)=\mathcal{N}\left(\boldsymbol{x}\mid\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}\left(\boldsymbol{\Lambda\mu}+\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b})\right),\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}\right)$

Proof:

$$\mathbb{E}[\boldsymbol{x}\mid\boldsymbol{y}]=\boldsymbol{\mu}+\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{A\mu}-\boldsymbol{b})=$$

$$=\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}\left(\boldsymbol{\Lambda\mu}+\boldsymbol{A}^T\boldsymbol{LA\mu}-\boldsymbol{A}^T\boldsymbol{LA\mu}+\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b})\right)$$

$$=\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}\left(\boldsymbol{\Lambda\mu}+\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{b})\right)$$

$$cov[\boldsymbol{x}\mid\boldsymbol{y}]=\left(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{LA}\right)^{-1}$$

Bayesian Inference for the Gaussian: Known Variance*

- Consider $X_1 \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$, *with prior* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. We want to infer μ with the variance σ^2 taken as known.

- Assuming a single data point $x_1 \sim \mathcal{N}(\mu, \sigma^2)$ we can derive:

$$\pi(\mu \mid x_1) \propto f(x_1 \mid \mu)\pi(\mu) \propto \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \Rightarrow$$

$$\pi(\mu \mid x_1) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) + \mu\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\mu-\mu_1)^2\right) \Rightarrow$$

$$\mu \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ with}$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}, \text{ and}$$

$$\mu_1 = \sigma_1^2\left(\frac{x_1}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

- The remaining of this lecture requires only elementary Bayesian inference calculations using Bayes' rule and appropriate prior distribution models. Bayesian inference will be discussed in detail in forthcoming lectures.

Bayesian Inference: Predictive distribution

- To predict the distribution of a new observation $X \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ in light of x_1, we use the predictive distribution as follows:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \mu)}_{Likelihood} \underbrace{\pi(\mu \mid x_1)}_{Posterior} d\mu \propto \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}} d\mu = \int e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right)} d\mu$$

- We can complete the square by treating the integrand above as a bivariate Gaussian in (x,μ). One can verify that:

$$\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right) = \frac{1}{2}\begin{pmatrix} x-\mu_1 & \mu-\mu_1 \end{pmatrix} \underbrace{\begin{pmatrix} \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & \frac{1}{\sigma^2}+\frac{1}{\sigma_1^2} \end{pmatrix}}_{\Sigma^{-1}} \begin{pmatrix} x-\mu_1 \\ \mu-\mu_1 \end{pmatrix} + const.$$

- From the above expression note that: $\Sigma = \begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}$

Bayesian Inference: Predictive distribution

- We derived in an earlier lecture that if we partition the mean and variance of a multivariate Gaussian as:

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

  then, the marginal

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a \mid \boldsymbol{\mu}_a, \Sigma_{aa})$$

- In our predictive distribution we need to integrate out μ. Thus based on the above result and $\mu = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}$, we have:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \mu)}_{\text{Likelihood}} \underbrace{\pi(\mu \mid x_1)}_{\text{Posterior}} d\mu = \mathcal{N}(x \mid \mu_1, \sigma^2 + \sigma_1^2)$$

- Note *the variance in the predictive distribution is the sum of model variance + variance of posterior uncertainty in μ.*

Bayesian Inference for the Gaussian

- Consider $\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(\mu, \sigma^2)$, *with prior* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

- The likelihood takes the form:

$$p(\boldsymbol{X} \mid \mu) = \prod_{n=1}^{N} f(x_n \mid \mu) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} \right)$$

- Note that in terms of μ this is not a probability density and is not normalized. Introducing the conjugate (Gaussian) prior on μ leads to:

$$\pi(\mu \mid \boldsymbol{X}) = \prod_{n=1}^{N} f(x_n \mid \mu)\pi(\mu) \propto \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right) \Rightarrow$$

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + \mu\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right) \propto \exp\left(-\frac{1}{2\sigma_N^2} (\mu - \mu_N)^2 \right)$$

Bayesian Inference for the Gaussian

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)+\mu\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$$

- So the posterior is a Gaussian as before with

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2} \Rightarrow \sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}, \text{ and}$$

$$\mu_N=\sigma_N^2\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)=\sigma_N^2\left(\frac{N\mu_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

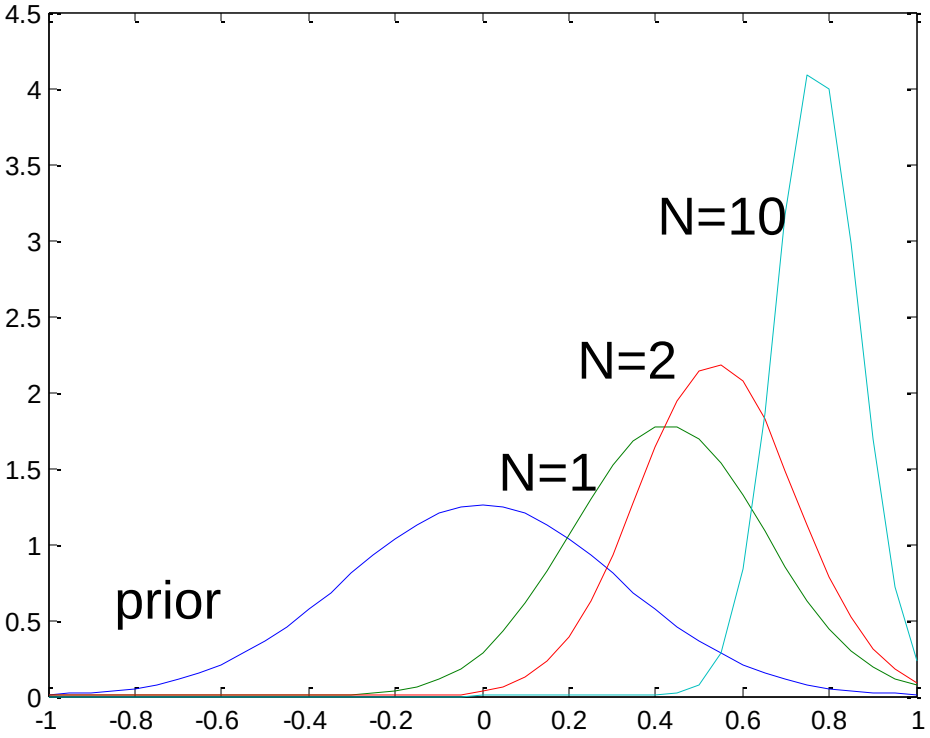
$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

- *The posterior precision is the sum of the precision of the prior plus one contribution of the data precision for each observed data point*.
- Observe the posterior mean for N→∞ and N→0.
- For N→∞ the posterior peaks around the μ_{ML} and the posterior variance goes to zero, i.e. the point MLE estimate is recovered within the Bayesian paradigm for infinite data.
- How about when $\sigma_0^2 \to \infty$? In this case note that $\sigma_N^2 \to \frac{\sigma^2}{N}$ *and* $\mu_N \to \mu_{ML}$

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \; with \; \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \; and \; \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

MatLab implementation



$$\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(0.8, 0.1), \; with \; prior \; \mu \sim \mathcal{N}(0, 0.1).$$

Sequential Bayesian Inference

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

- We can easily derive sequential estimates of the posterior variance and mean.

They are as follows:

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}, \text{ and } \mu_N = \frac{\sigma_N^2}{\sigma_{N-1}^2} \mu_{N-1} + \frac{\sigma_N^2}{\sigma^2} x_N$$

Example of Linear Gaussian Systems: Inferring the Mean

- We revisit this Bayesian inference problem for the Gaussian. Consider

$$\boldsymbol{y} = \{y_1, y_2, ..., y_N\} \sim \mathcal{N}(y \mid x, \sigma^2 = \lambda_y^{-1}), \text{ with prior } x \sim \mathcal{N}(x \mid \mu_0, \sigma_0^2 = \lambda_0^{-1}).$$

- To put the likelihood for the N-data set in the form of a linear Gaussian system, let:

$$p(\boldsymbol{y} \mid x) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}x + \boldsymbol{b}, \boldsymbol{\Sigma}_y\right), \boldsymbol{A} = \mathbf{1}_N \text{ (column vector of 1's)}, \boldsymbol{b} = \mathbf{0}, \boldsymbol{\Sigma}_y^{-1} = diag(\lambda_y \boldsymbol{I})$$

- Then from our conditional results given earlier:

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x} \mid \left(\Lambda + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A}\right)^{-1}\left(\Lambda \boldsymbol{\mu} + \boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})\right), \left(\Lambda + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A}\right)^{-1}\right)$$

$$p(x \mid \boldsymbol{y}) = \mathcal{N}\left(x \mid \left(\lambda_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I} \mathbf{1}_N\right)^{-1}\left(\lambda_0 \mu_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I}(\boldsymbol{y} - \boldsymbol{0})\right), \left(\lambda_0 + \mathbf{1}_N^T \lambda_y \boldsymbol{I} \mathbf{1}_N\right)^{-1}\right)$$

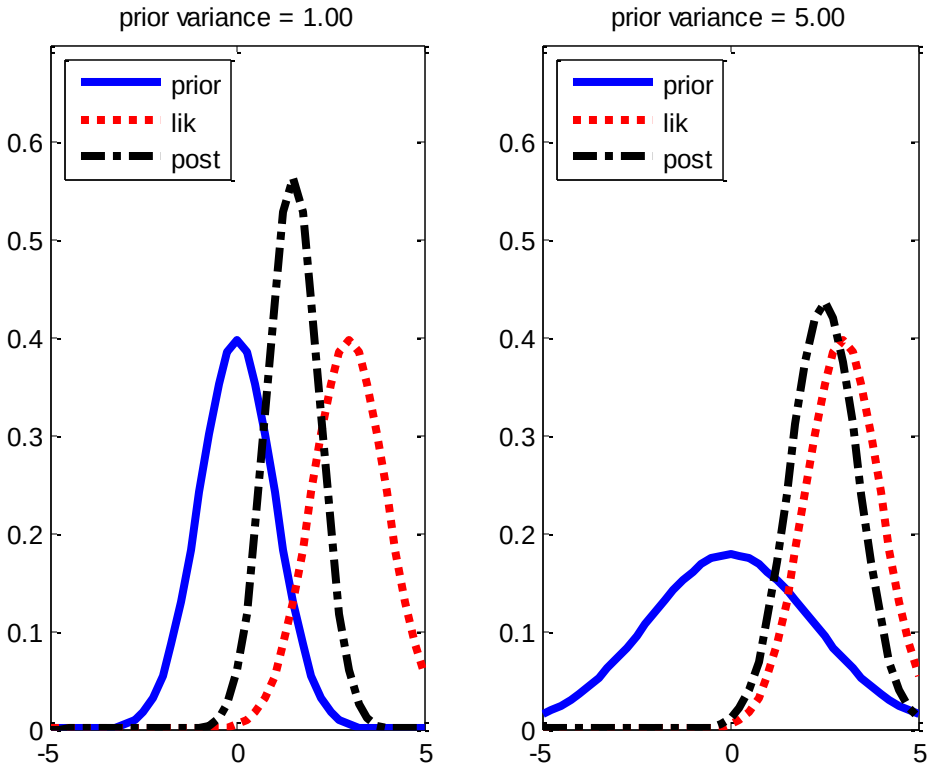
- This can be simplified as $p(x|y) = \mathcal{N}(x|\mu_N, \lambda_N^{-1})$ where:

$$p(x|y) = \mathcal{N}\left(x|\frac{N\lambda_y}{\lambda_0 + N\lambda_y}\bar{y} + \frac{\lambda_0}{\lambda_0 + N\lambda_y}\mu_0, \left(\lambda_0 + N\lambda_y\right)^{-1}\right)$$

- The precision is the prior precision + N measurement precisions. The mean is the weighted average of the MLE and prior mean.

Example of Linear Gaussian Systems: Inferring the Mean

gaussInferParamsMean1d
from PMTK



- Inference about x given a single noisy observation $y = 3$.
  - (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0.
  - (b) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE

Example of Linear Gaussian Systems: Inferring the Mean

- We can re-write these results in terms of variances recovering the results we have seen earlier in this lecture.

$$x \mid \mathbf{Y} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and}$$

$$\mu_N = \sigma_N^2 \left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \sigma_N^2 \left(\frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

Shrinkage and Signal-To-Noise Ratio

$$\sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

- The posterior precision is the sum of the precision of the prior plus one contribution of the data precision for each observed data point. For N→∞ the posterior peaks around the μ_{ML} and the posterior variance goes to zero, i.e. MLE estimate is recovered within the Bayesian paradigm.

- If we apply the data sequentially, we can write for the posterior mean *after the collection of one data point (N=1), i.e.* $\mu_{ML} = y$ the following:

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma_0^2 + \sigma^2} \; (\textit{shrinkage of the data } y \textit{ towards the prior mean } \mu_0)$$

- Shrinkage is often measured also with the *signal-to-noise ratio*:

$$SNR = \frac{\mathbb{E}\left[X^2\right]}{\mathbb{E}\left[\varepsilon^2\right]} = \frac{\sigma_0^2 + \mu_0^2}{\sigma^2}, \; \textit{for } y = x + \varepsilon \; (\textit{observed signal}), \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- How about when $\sigma_0^2 \to \infty$? In this case note that $\sigma_N^2 \to \frac{\sigma^2}{N}$ *and* $\mu_N \to \mu_{ML}$

Estimating the Mean of a Multivariate Gaussian

- Consider the following linear Gaussian system:

$$\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma}_y), \text{ with prior } \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Consider an effective observation $\bar{\boldsymbol{y}}|\mathbf{x} \sim \mathcal{N}(\bar{\boldsymbol{y}}|\boldsymbol{x}, \frac{1}{N}\boldsymbol{\Sigma}_y)$.*

- From our earlier results with **A=I**, we have:

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x} \mid \left(\Lambda + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\right)^{-1}\left(\Lambda\boldsymbol{\mu} + \boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})\right), \left(\Lambda + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A}\right)^{-1}\right)$$

$$p(\boldsymbol{x}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N) = \mathcal{N}\left(\boldsymbol{x}|\left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}_y^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + N\boldsymbol{\Sigma}_y^{-1}\bar{\boldsymbol{y}}\right), \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}_y^{-1}\right)^{-1}\right)$$

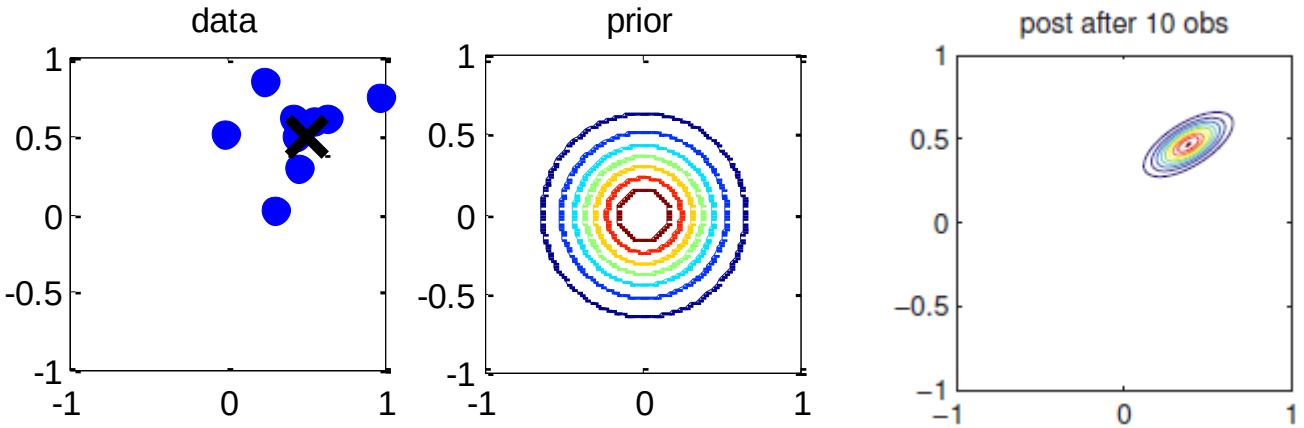
---

* Note that for $\boldsymbol{y} \underset{N}{=} \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma}_y)$, the term in the exponential of the likelihood is $-\frac{1}{2}\sum_{i=1}(\boldsymbol{y}_i - \boldsymbol{x})^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}_i - \boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T(N\boldsymbol{\Sigma}_y^{-1})\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}_y^{-1}N\bar{\boldsymbol{y}} + const$. This completing the square gives $\mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{y}}, \frac{1}{N}\boldsymbol{\Sigma}_y)$ or equivalently a likelihood for the effective observation $p(\bar{\boldsymbol{y}}|\boldsymbol{x}) = \mathcal{N}\left(\bar{\boldsymbol{y}}|\boldsymbol{x}, \frac{1}{N}\boldsymbol{\Sigma}_y\right)$

Bayesian Inference for the Mean of a 2d Gaussian

- Bayesian inference for the mean of a 2d Gaussian. (a) Data generated as $\mathbf{y}_i \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}_y)$, $\mathbf{x} = [0.5, 0.5]^T$ and $\mathbf{\Sigma}_y = 0.1[2, 1; 1, 1]$. $\mathbf{\Sigma}_y$ is known but $\mathbf{x}$ is unknown. The black cross represents $\mathbf{x}$. (b) The prior is $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, 0.1\mathbf{I}_2)$. (c) The posterior is computed after 10 data points.

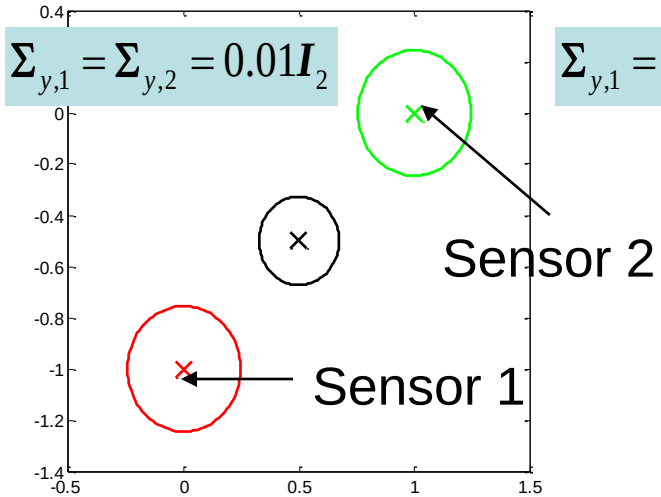
- Think of this as identifying a missile location ***x*** from noisy measurements $\mathbf{y}_i$!

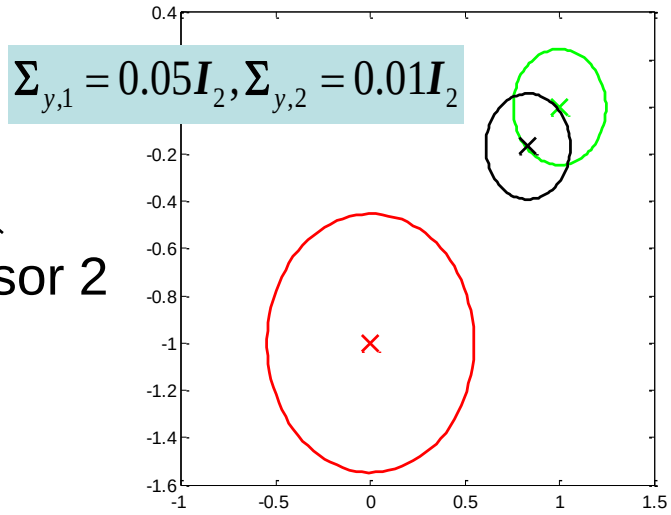


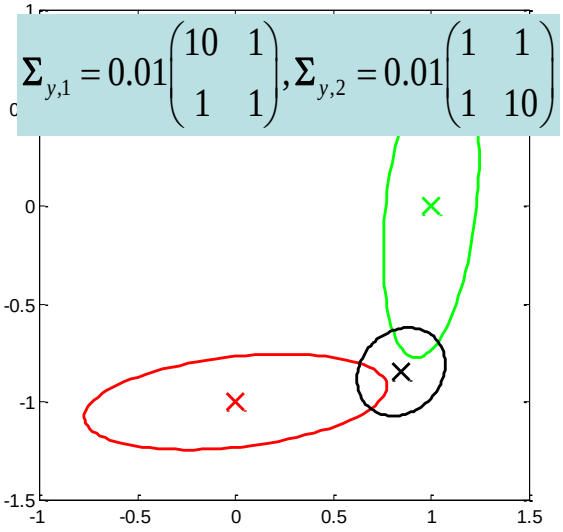
gaussInferParamsMean2d
from PMTK

Sensor Fusion

- We observe $\mathbf{y}_1 = (0, -1)$ (red cross) and $\mathbf{y}_2 = (1, 0)$ (green cross) and infer $\mathbb{E}(\boldsymbol{\mu}|\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{\theta})$ (black cross). (a) Equally reliable sensors, so the posterior mean estimate is in between the two circles. (b) Sensor 2 is more reliable, so the estimate shifts more towards the green circle. (c) Sensor 1 is more reliable in the vertical direction, Sensor 2 is more reliable in the horizontal direction. The estimate is an appropriate combination of the two measurements.







$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}_0 = \mathbf{0}, \Sigma_0 = 10^{10} \boldsymbol{I}_2\right)$$

$$\boldsymbol{y}_1 \sim \mathcal{N}\left(\boldsymbol{x}, \Sigma_{y,1}\right), \boldsymbol{y}_2 \sim \mathcal{N}\left(\boldsymbol{x}, \Sigma_{y,2}\right)$$

sensorFusion2d from PMTK

Interpolating Noisy Data

- We revisit the 1D interpolation example from an earlier lecture but now with noisy data. We collect N observations y_i of $x_1,\ldots,x_N$.

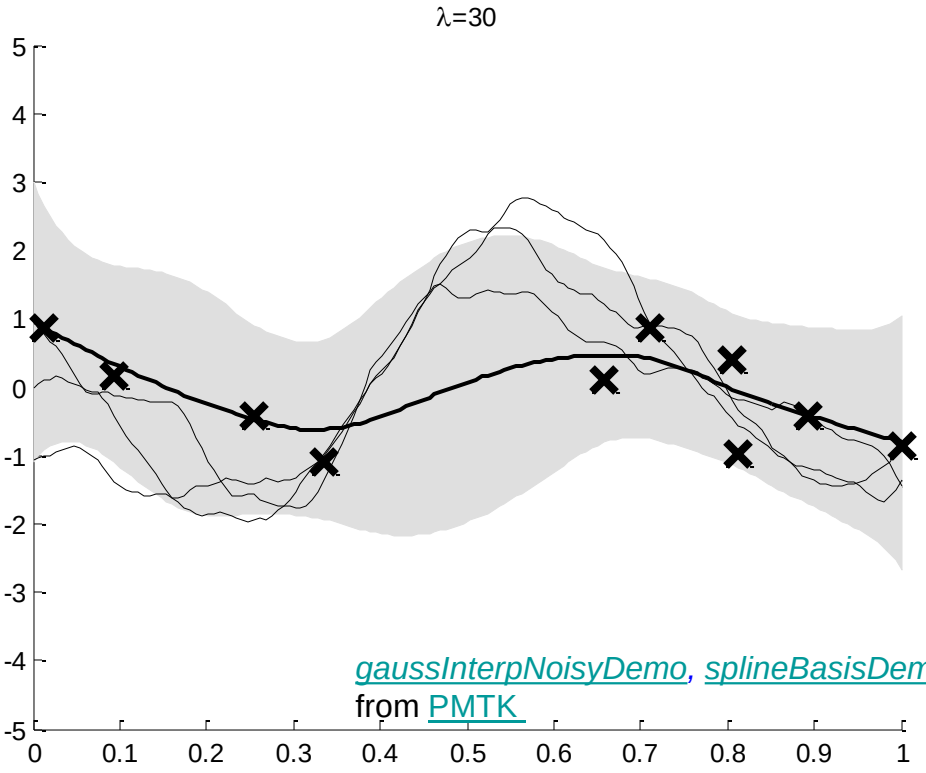
$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_y\right), \boldsymbol{\Sigma}_y = \sigma^2 \boldsymbol{I}, e.g.\ \boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}_{N=2\times D=4}$$

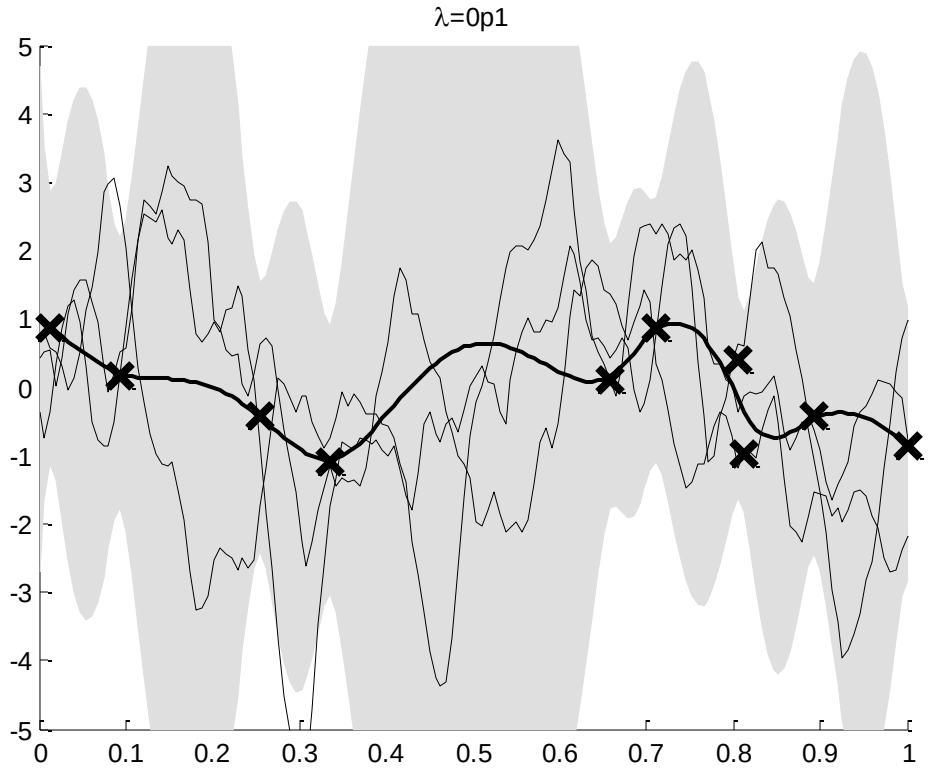
- Here A is an N×D matrix that picks the observed elements out of $\boldsymbol{x}$.
- The prior is as before: $p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{0}, \lambda\left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}\right), \boldsymbol{\Sigma}_x = \lambda\left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}$
- Using the linear Gaussian system model, we can compute the needed posterior $p(\boldsymbol{x}\,|\,\boldsymbol{y})$ as before (an example on the next slide).
- The posterior mean can also be computed by solving the following (regularized) optimization problem (the 2nd term penalizes rapid variability of the data – 1st derivative Tikhonov regularization)

$$\min_{x_i} \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{D}\left[\left(x_j - x_{j-1}\right)^2 + \left(x_j - x_{j+1}\right)^2\right], x_0 = x_1, x_{D+1} = x_D$$

  - *D. Calvetti and E. Somersalo, Introduction to Bayesian Scientific Computing, 2007*

Interpolating Noisy Data





gaussInterpNoisyDemo, *splineBasisDemo* from PMTK

- *We now see that the prior precision λ effects the posterior mean as well as the posterior variance* (in comparison to the case of no noise)
- For a strong prior (large λ), the estimate is very smooth, and the uncertainty is low but for a weak prior (small λ), the estimate is wiggly, and the uncertainty (away from the data) is high.