Bayesian Methods for Machine Learning

Lecture 1 - Fundamentals of Bayesian modeling and inference

Simon Leglaive

CentraleSupélec

Introductory example

The following example and drawings are adapted from a tutorial on Bayesian Learning for Signal Processing given by Antoine Deleforge at the LVA/ICA 2015 Summer School.

Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

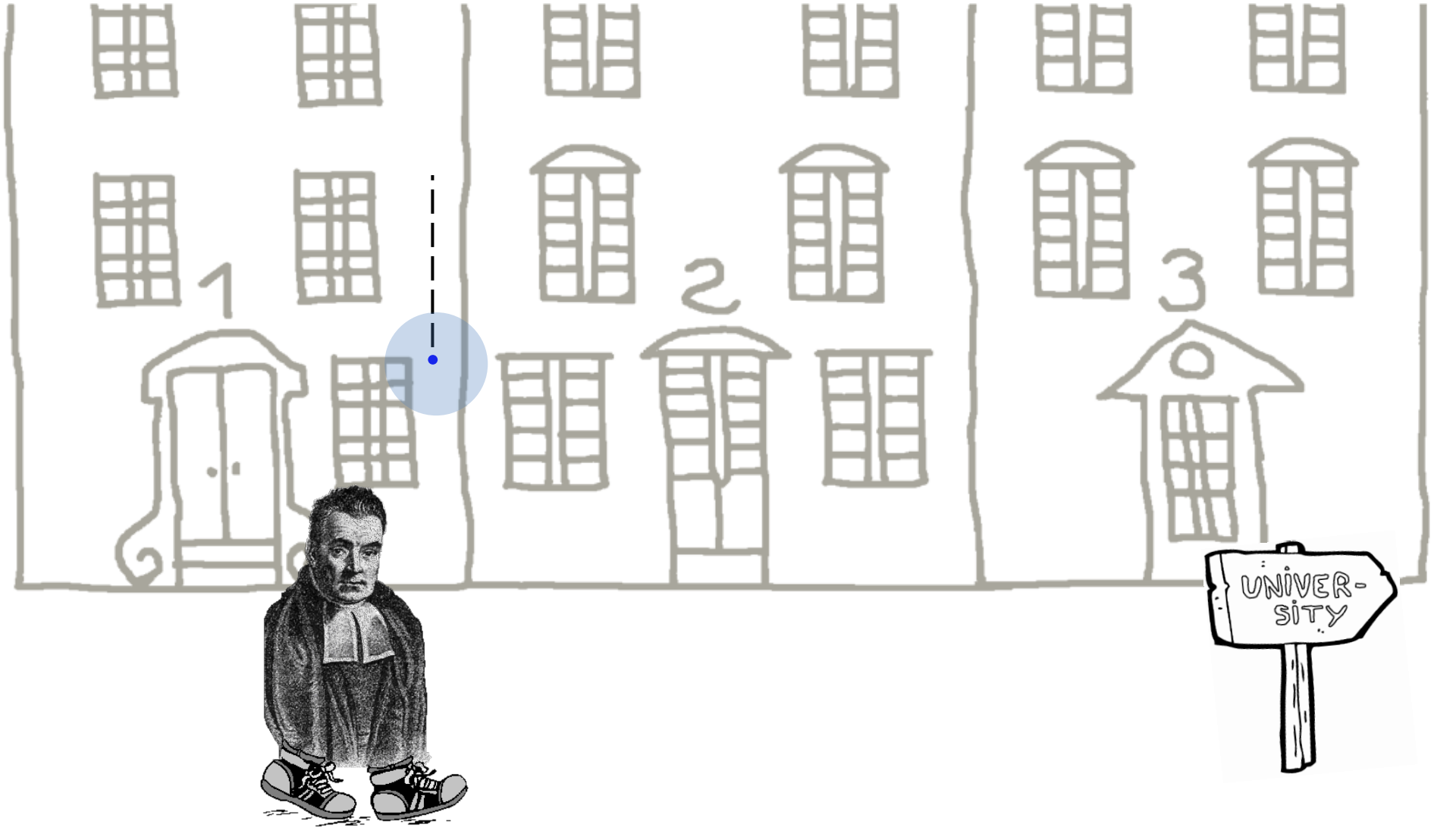


Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

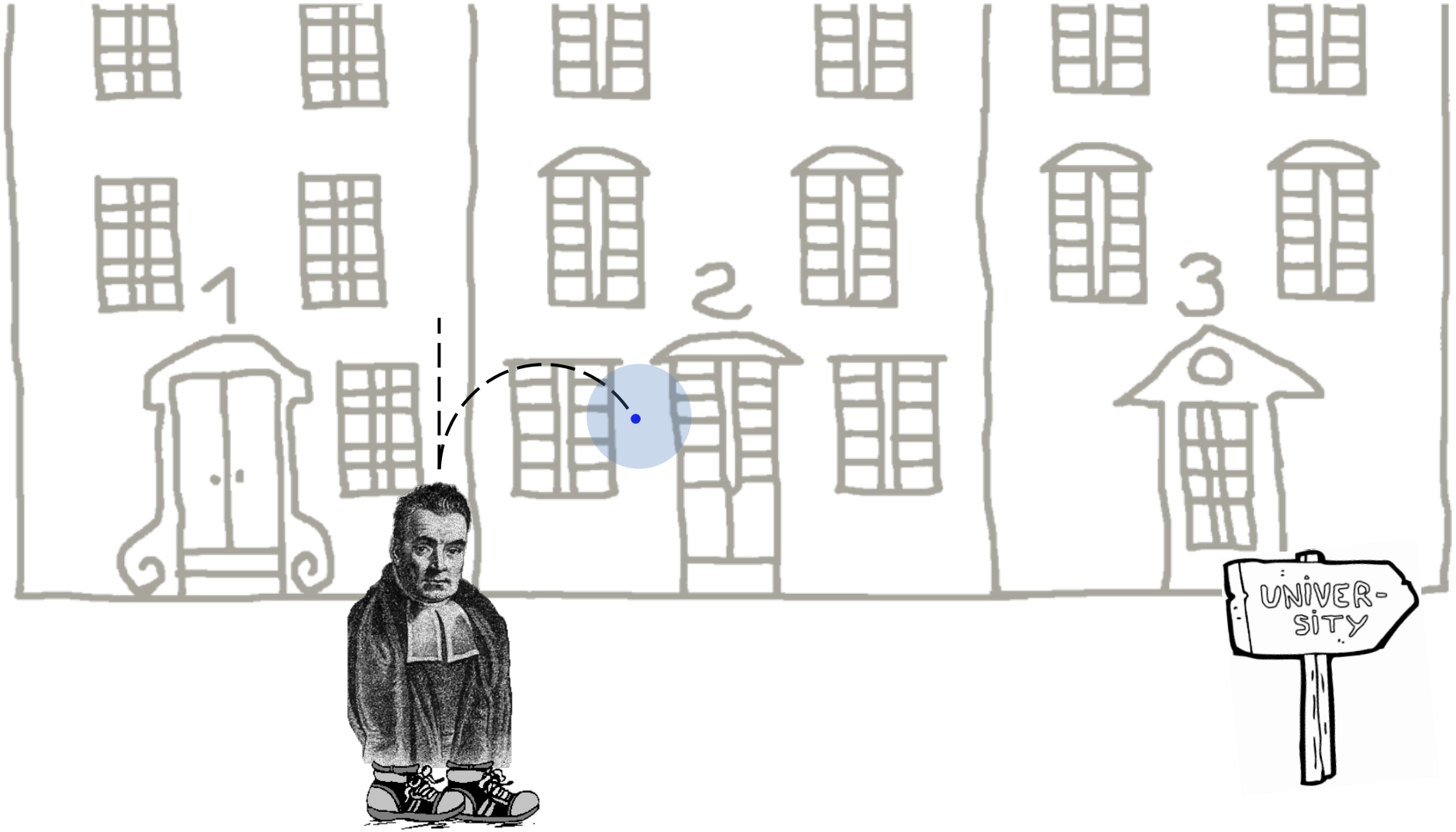


Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

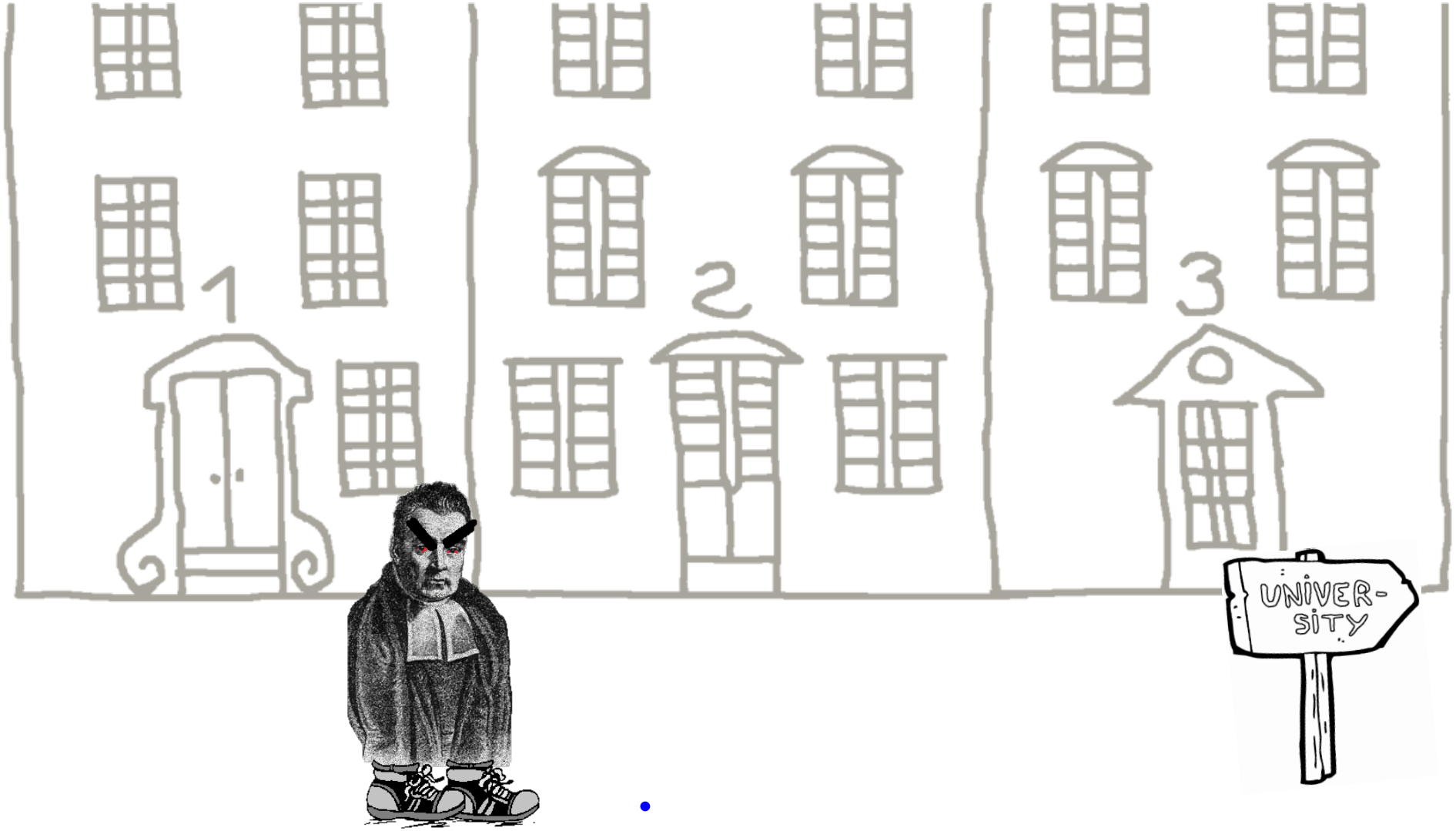


Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

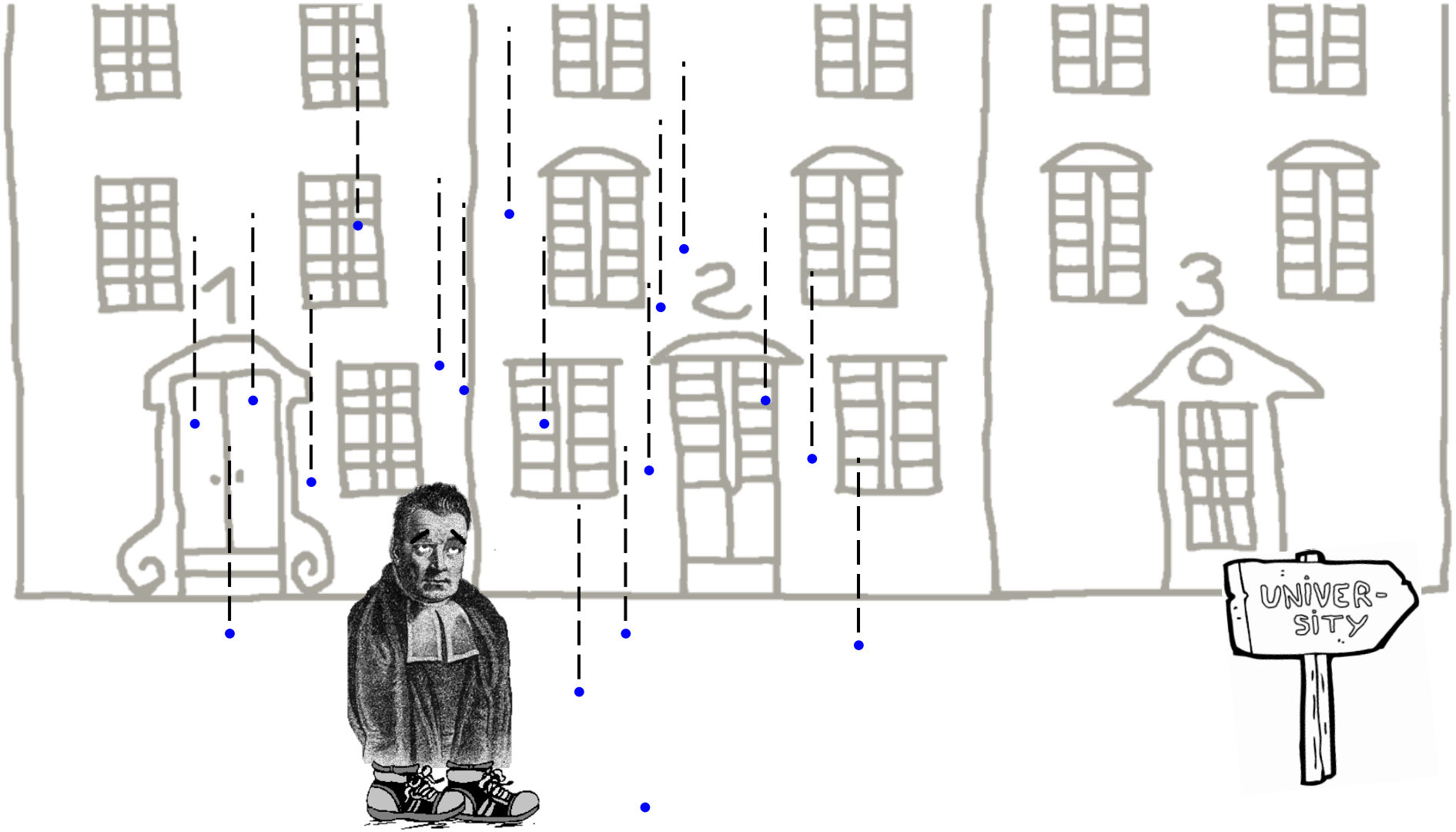


Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

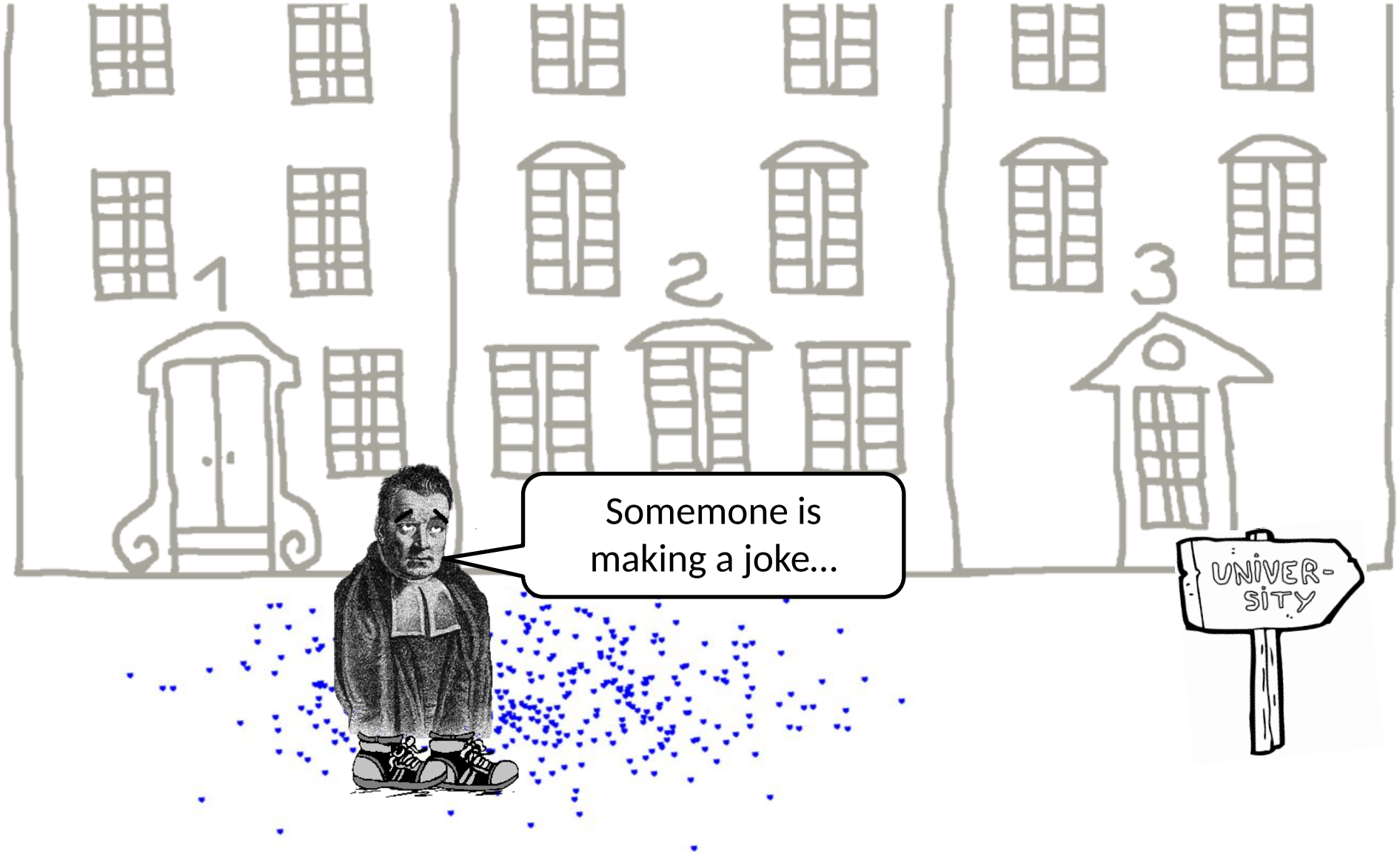
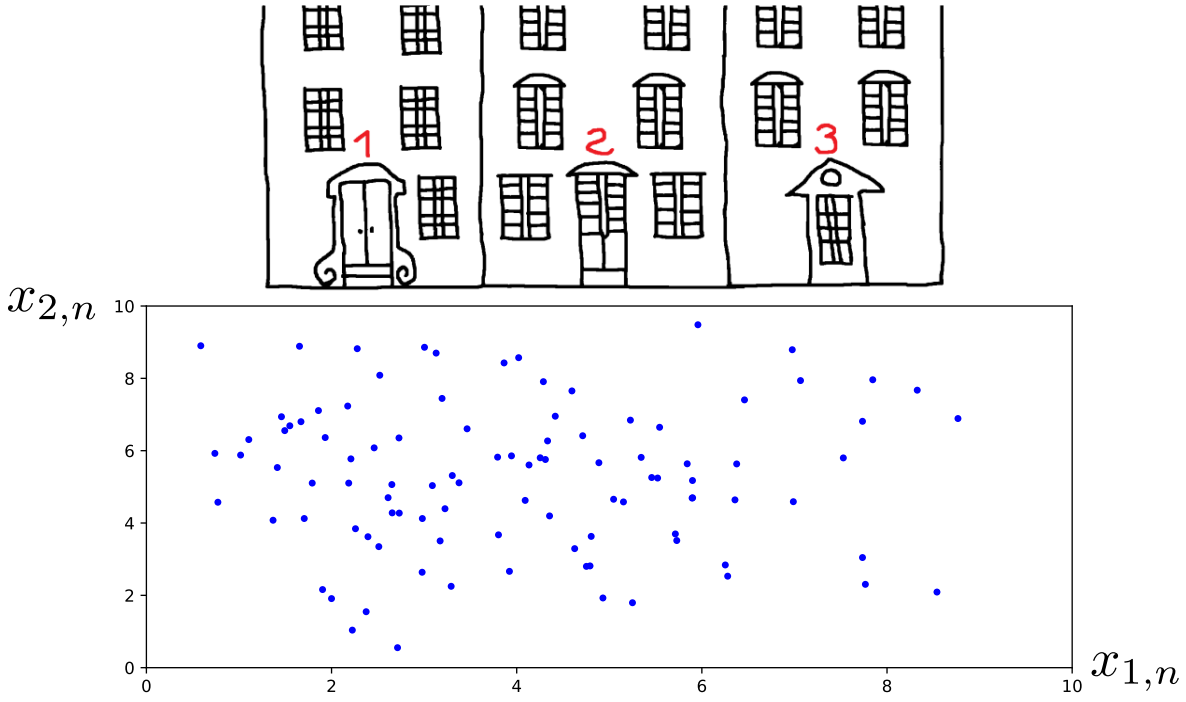


Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

Image credit: Antoine Deleforge, Tutorial on Bayesian Learning for Signal Processing, LVA/ICA 2015 Summer School

Modeling



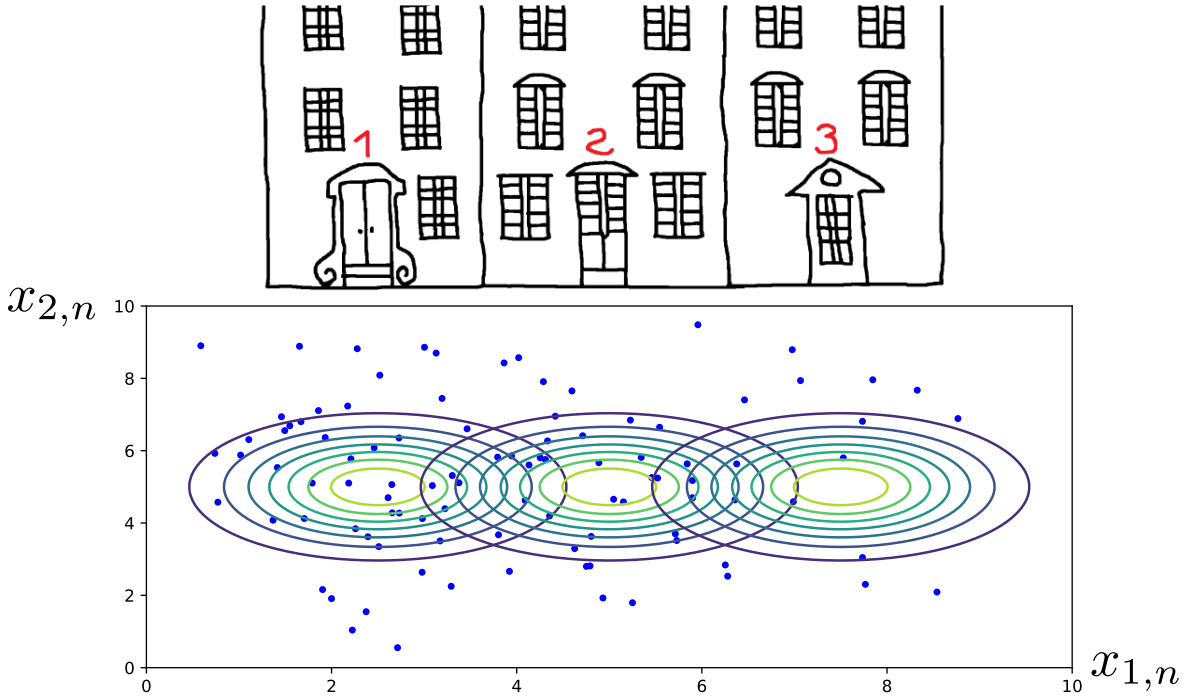
Observed variables

$$\{\mathbf{x}_n = [x_{1,n}, x_{2,n}]^\top \in \mathbf{R}^2\}_{n=1}^N$$

Latent (hidden) variable

$$z \in \{1, 2, 3\}$$

Modeling



Observed variables

$$\{\mathbf{x}_n = [x_{1,n}, x_{2,n}]^\top \in \mathrm{R}^2\}_{n=1}^N$$

Latent (hidden) variable

$$z \in \{1, 2, 3\}$$

Observation model

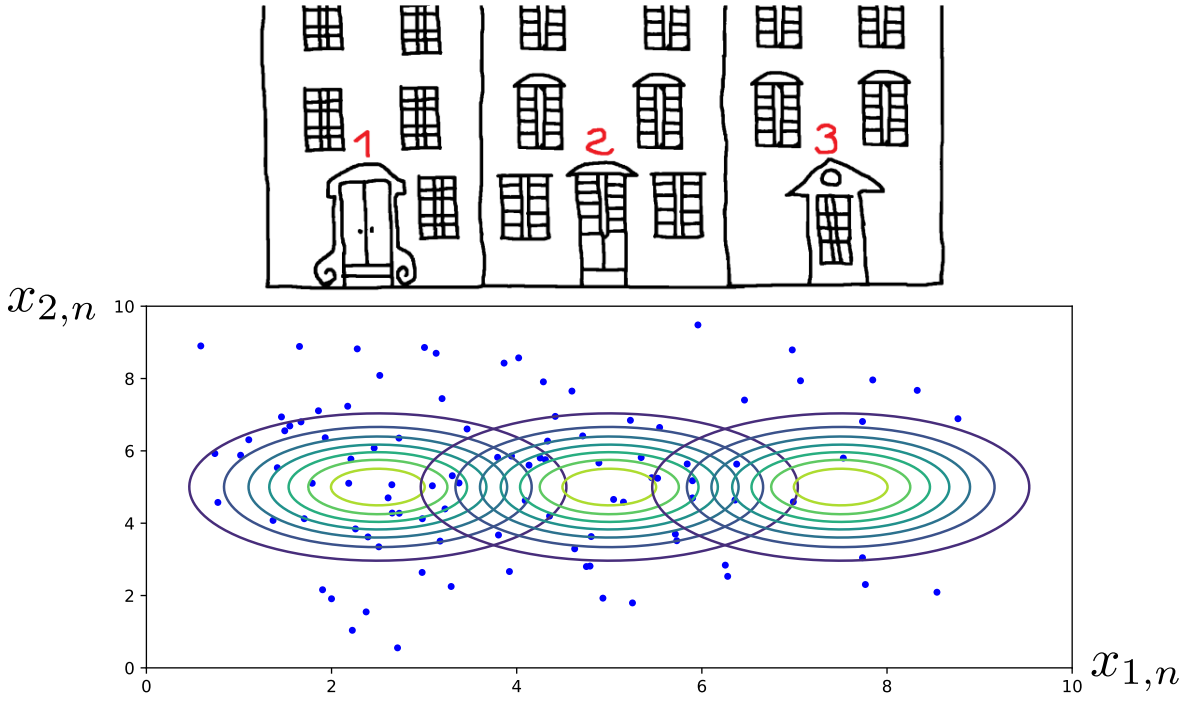
- $p(\mathbf{x}_n|z=1) = \mathrm{N}(\mathbf{x}_n; \boldsymbol{\mu}_1, \sigma^2\mathbf{I})$
- $p(\mathbf{x}_n|z=2) = \mathrm{N}(\mathbf{x}_n; \boldsymbol{\mu}_2, \sigma^2\mathbf{I})$
- $p(\mathbf{x}_n|z=3) = \mathrm{N}(\mathbf{x}_n; \boldsymbol{\mu}_3, \sigma^2\mathbf{I})$

with $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ and σ^2 known and fixed. All observations are assumed to be independent and identically distributed (i.i.d.) given the latent variable z.

Maximum likelihood (ML) estimation

The probability density function of the multivariate Gaussian distribution is defined by:

$$\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$



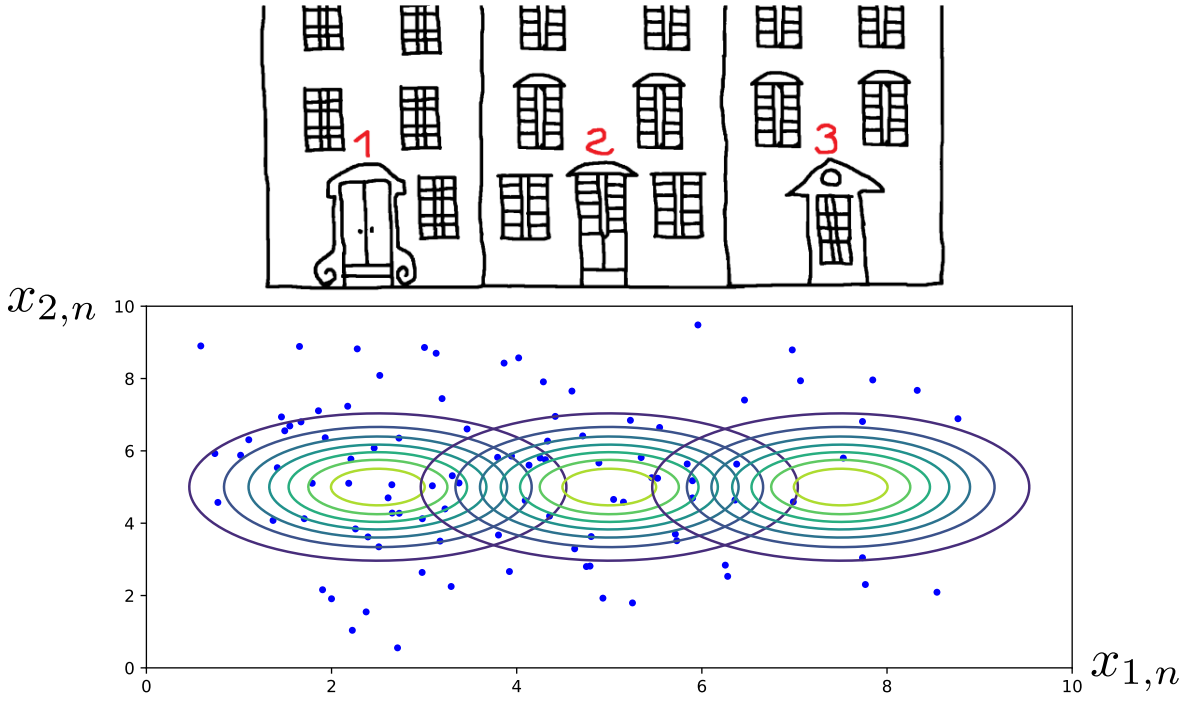
Likelihood

$$\begin{aligned}\mathrm{L}(i) &= p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N|z=i)\\ &= \prod_{n=1}^{N} p(\mathbf{x}_n|z=i)\\ &= \prod_{n=1}^{N} \mathrm{N}\left(\mathbf{x}_n;\boldsymbol{\mu}_i,\sigma^2\mathbf{I}\right)\end{aligned}$$

Maximum likelihood (ML) estimation

The probability density function of the multivariate Gaussian distribution is defined by:

$$\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$



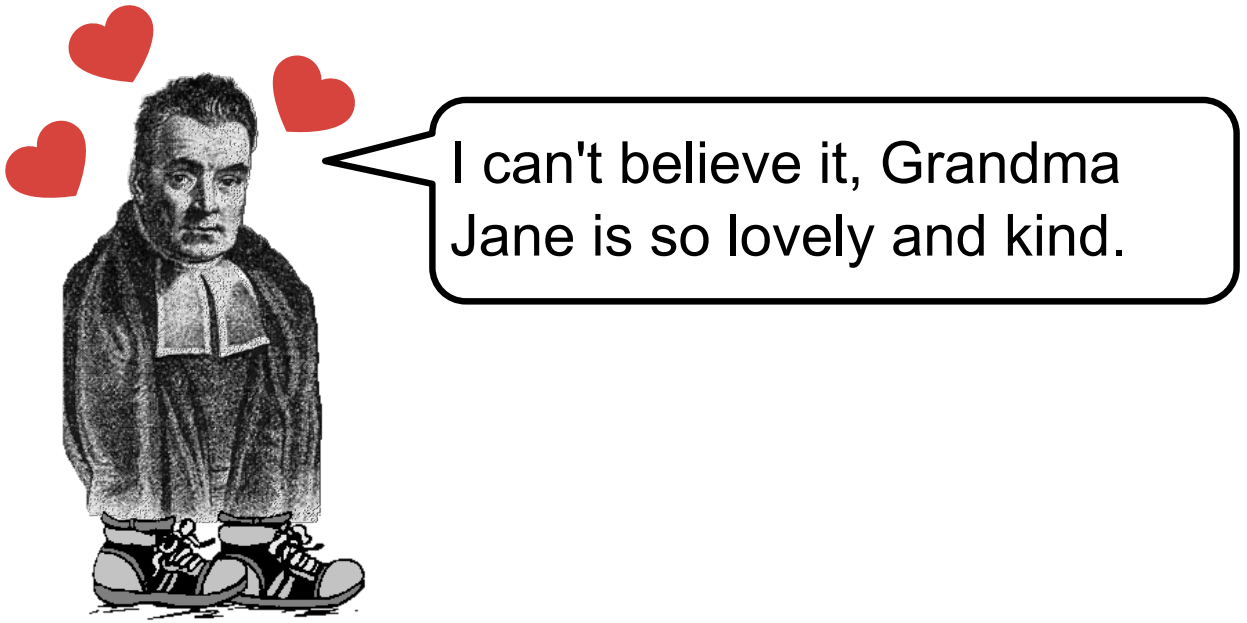
Log-likelihood

$$\begin{aligned}\mathrm{L}(i) &= \ln p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N | z = i) \\ &= \ln \prod_{n=1}^{N} p(\mathbf{x}_n | z = i) \\ &= \sum_{n=1}^{N} \ln \mathrm{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_i, \sigma^2\mathbf{I}\right)\end{aligned}$$

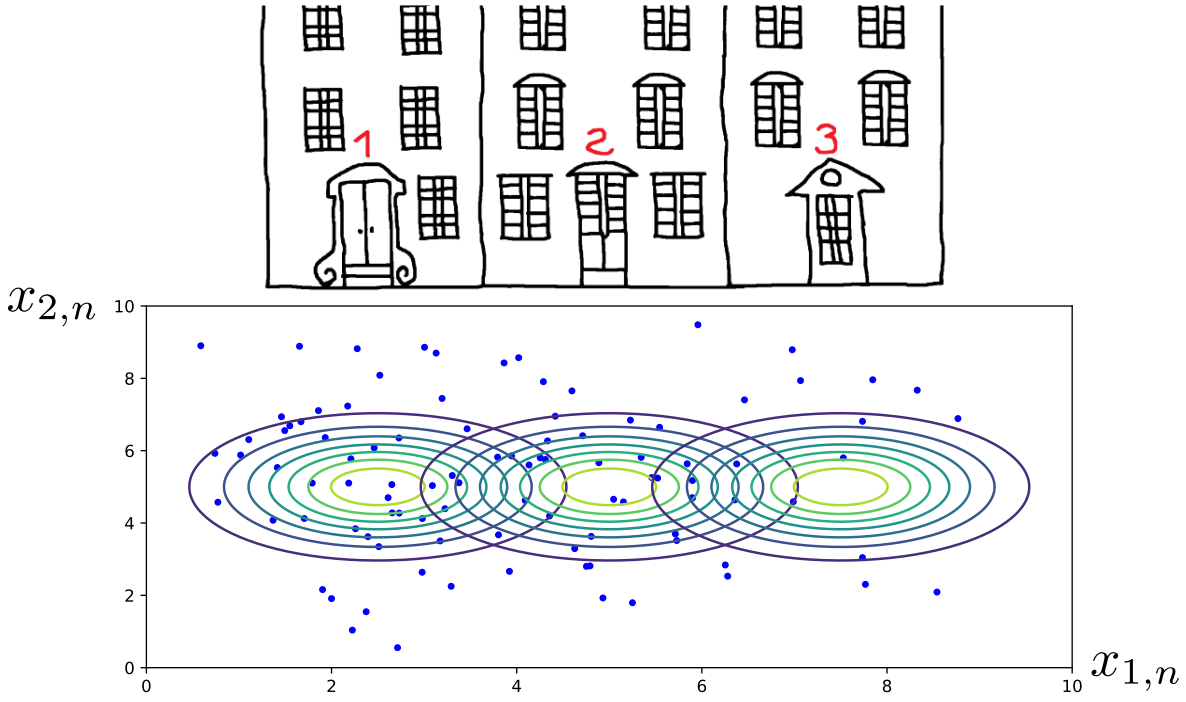
```python
import numpy as np
from scipy.stats import multivariate_normal

mu_1 = np.array([2.5, 5])
mu_2 = np.array([5, 5])
mu_3 = np.array([7.5, 5])

Sigma = 100*np.eye(2)

log_likelihood = np.zeros(3)

for i, mu in enumerate([mu_1, mu_2, mu_3]):
    log_likelihood[i] = np.sum(multivariate_normal.logpdf(data, mean=mu, cov=Sigma))

print(log_likelihood)
```

```
>>> [-649.56488429 -648.88937372 -654.46386315]
```

$$\hat{z} = \underset{i \in \{1,2,3\}}{\arg\max}\, \mathrm{L}(i) = 2$$


I can't believe it, Grandma Jane is so lovely and kind.

Bayesian modeling



Observed variables

$$\{\mathbf{x}_n = [x_{1,n}, x_{2,n}]^\top \in \mathrm{R}^2\}_{n=1}^N$$

Latent (hidden) variable

$$z \in \{1, 2, 3\}$$

Observation model (likelihood)

$$p(\mathbf{x}_n | z = i) = \mathrm{N}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I})$$

Prior

- $p(z = 1) = 0.3$ (student house)
- $p(z = 2) = 0.1$ (grandma Jane)
- $p(z = 3) = 0.6$ (family with kids)

Bayesian inference

Bayes' Theorem allows us to compute the posterior distribution given the likelihood and the priors:

$$
\begin{aligned}
p(z=i|\mathbf{x}_1,...,\mathbf{x}_N) &= \frac{p(\mathbf{x}_1,...,\mathbf{x}_N|z=i)p(z=i)}{p(\mathbf{x}_1,...,\mathbf{x}_N)} \\
&= \frac{p(\mathbf{x}_1,...,\mathbf{x}_N|z=i)p(z=i)}{\sum_{k=1}^{3} p(\mathbf{x}_1,...,\mathbf{x}_N,z=k)} && \text{(using the sum rule)} \\
&= \frac{p(\mathbf{x}_1,...,\mathbf{x}_N|z=i)p(z=i)}{\sum_{k=1}^{3} p(\mathbf{x}_1,...,\mathbf{x}_N|z=k)p(z=k)} && \text{(using the product rule)} \\
&= \frac{p(z=i)\prod_n p(\mathbf{x}_n|z=i)}{\sum_{k=1}^{3} p(z=k)\prod_n p(\mathbf{x}_n|z=k)}
\end{aligned}
$$

```
# z=1 : student house, z=2: grandma Jane, z=3 family with kids

prior = np.array([0.3, 0.1, 0.6])
log_prior = np.log(prior)
log_joint = log_likelihood + log_prior

print(log_prior)
print(log_likelihood)
print(log_joint)

>>>[-1.2039728  -2.30258509 -0.51082562]
>>>[-649.56488429 -648.88937372 -654.46386315]
>>>[-650.7688571  -651.19195881 -654.97468877]

log_marginal = np.log(np.sum(np.exp(log_joint)))
log_posterior = log_joint - log_marginal
posterior = np.exp(log_posterior)
```

Decision

The posterior contains all the information about the latent variable we care about, but it does not directly tell Bayes which house is the guilty one.

From the posterior $p(z = i|\mathbf{x}_1, ..., \mathbf{x}_N), i \in \{1, 2, 3\}$, Bayes needs to choose a single value $\hat{z}$ which will serve as a point estimate of the latent variable z.

In other words, he needs to take a decision.

To do so, Bayes needs to build a loss function $l(\hat{z} = k, z = i)$ which tells "how bad" would it be to decide $\hat{z} = k$ if the "true value" of the latent variable was $z = i$, with $(k, i) \in \{1, 2, 3\}^2$.

- Of course he sets $l(\hat{z} = k, z = k) = 0$ for all k.
- Bayes appreciates grandma Jane a lot, he really doesn't want to accuse her by mistake, so he sets $l(\hat{z} = 2, z \neq 2) = 10$.
- On the contrary, he does not really care if he accuses the kids or the students by mistake, so he sets $l(\hat{z} = 1, z \neq 1) = l(\hat{z} = 3, z \neq 3) = 1$.

Loss function	students ($\hat{z} = 1$)	grandma Jane ($\hat{z} = 2$)	kids ($\hat{z} = 3$)
students ($z = 1$)	0	10	1
grandma Jane ($z = 2$)	1	0	1
kids ($z = 3$)	1	10	0

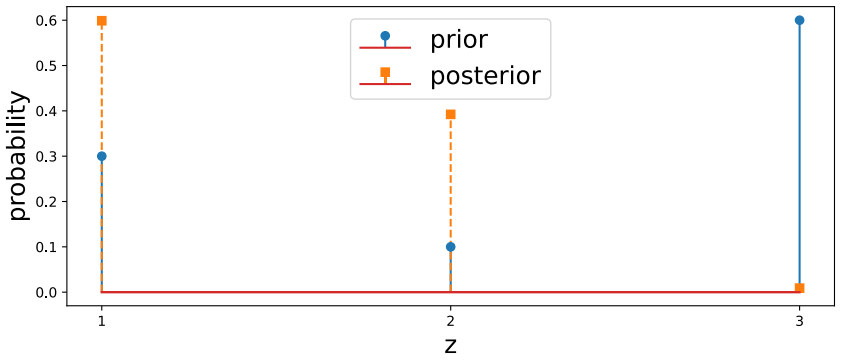
The posterior expected loss is defined as the expectation of the loss taken with respect to the posterior distribution:

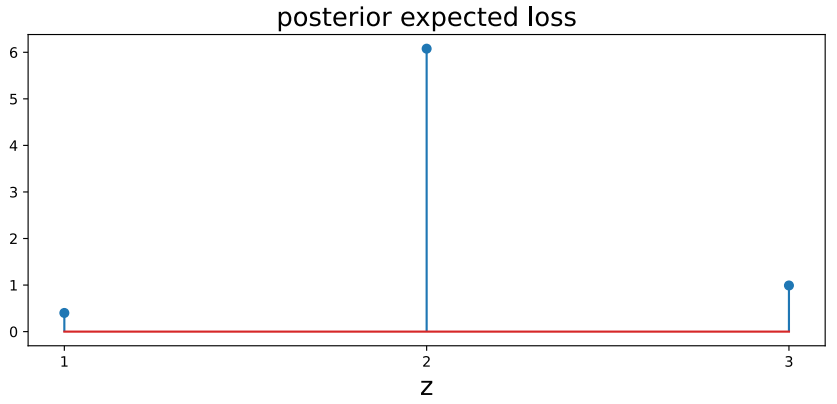
$$\begin{aligned}\mathrm{L}(\hat{z}=k) &= \mathrm{E}_{p(z=i|\mathbf{x}_1,\dots,\mathbf{x}_N)}[l(\hat{z}=k,z=i)] \\ &= \sum_{i=1}^{3} l(\hat{z}=k,z=i)p(z=i|\mathbf{x}_1,\dots,\mathbf{x}_N).\end{aligned}$$

It consists in weighting the loss using the knowledge/uncertainty that Bayes has at the moment he wants to take the decision, which is of course carried by the posterior.

To take his decision, Bayes finally minimizes the loss:

$$\hat{z}_{\text{guilty}} = \arg\min_{k\in\{1,2,3\}} \mathrm{L}(\hat{z}=k).$$



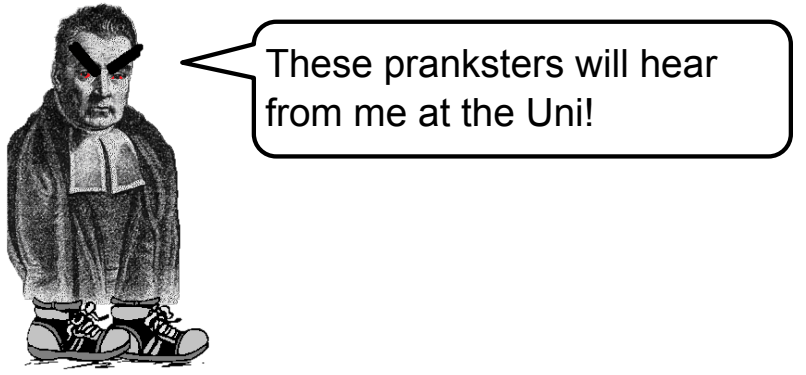


The decision that minimizes the posterior expected loss is $\hat{z}_{\text{guilty}} = 1$ (the students!).

It also corresponds to the maximum a posteriori (MAP) estimate:

$$\hat{z}_{\text{MAP}} = \underset{i \in \{1,2,3\}}{\arg\max}\, p(z = i | \mathbf{x}_1, ..., \mathbf{x}_N).$$

But we can see that "the ranking" between the suspects is different according to the posterior and the posterior expected loss.

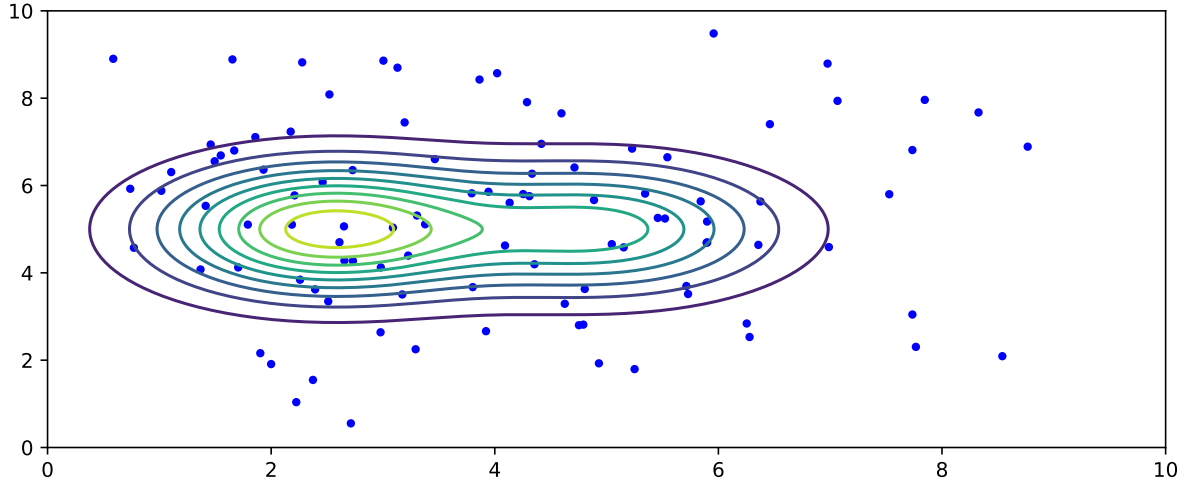


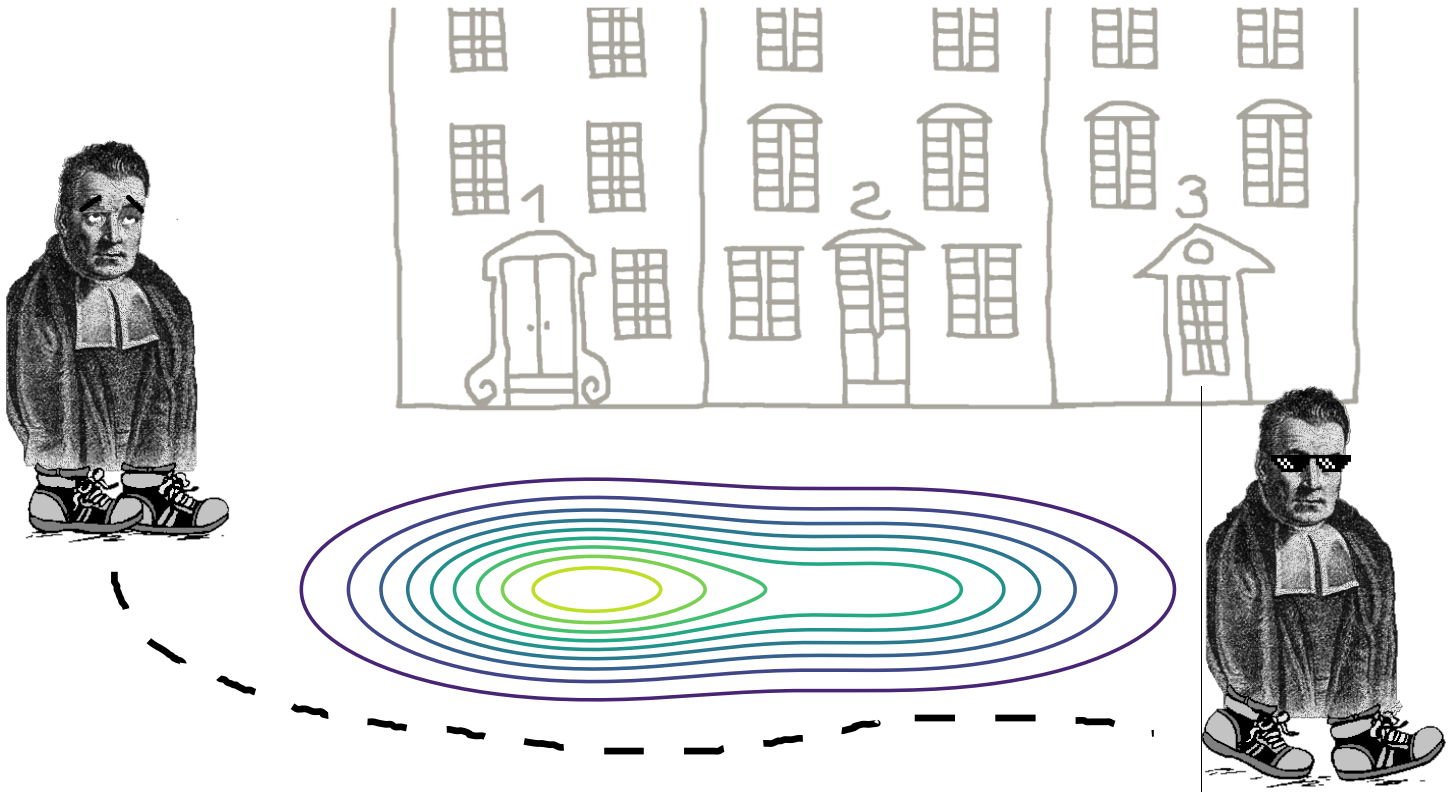
Prediction

The next day, Bayes goes to the university armed with its predictive posterior:

$$
\begin{aligned}
p(\mathbf{x}_{\text{new}}|\mathbf{x}_1, ..., \mathbf{x}_N) &= \sum_{i=1}^{3} p(\mathbf{x}_{\text{new}}, z = i|\mathbf{x}_1, ..., \mathbf{x}_N) \\
&= \sum_{i=1}^{3} p(\mathbf{x}_{\text{new}}|z = i, \mathbf{x}_1, ..., \mathbf{x}_N) p(z = i|\mathbf{x}_1, ..., \mathbf{x}_N) \\
&= \sum_{i=1}^{3} p(\mathbf{x}_{\text{new}}|z = i) p(z = i|\mathbf{x}_1, ..., \mathbf{x}_N) \\
&= \sum_{i=1}^{3} \mathrm{N}\left(\mathbf{x}_{\text{new}}; \boldsymbol{\mu}_i, \sigma^2\mathbf{I}\right) p(z = i|\mathbf{x}_1, ..., \mathbf{x}_N) \\
&= \mathrm{E}_{p(z=i|\mathbf{x}_1,...,\mathbf{x}_N)}\left[\mathrm{N}\left(\mathbf{x}_{\text{new}}; \boldsymbol{\mu}_i, \sigma^2\mathbf{I}\right)\right]
\end{aligned}
$$

```
x, y = np.mgrid[0:10:.01, 0:10:.01]
pos = np.dstack((x, y))

pred_1 = multivariate_normal(mean=mu_1, cov=Sigma_viz).pdf(pos)
pred_2 = multivariate_normal(mean=mu_2, cov=Sigma_viz).pdf(pos)
pred_3 = multivariate_normal(mean=mu_3, cov=Sigma_viz).pdf(pos)

pred = pred_1*posterior[0] + pred_2*posterior[1] + pred_3*posterior[2]


fig3 = plt.figure(figsize=(10,4))
ax3 = fig3.add_subplot(111)
ax3.scatter(data[:,0], data[:,1], color='blue', s=10)
plt.xlim(0, 10)
plt.ylim(0, 10)

ax3.contour(x, y, pred, 10)
```



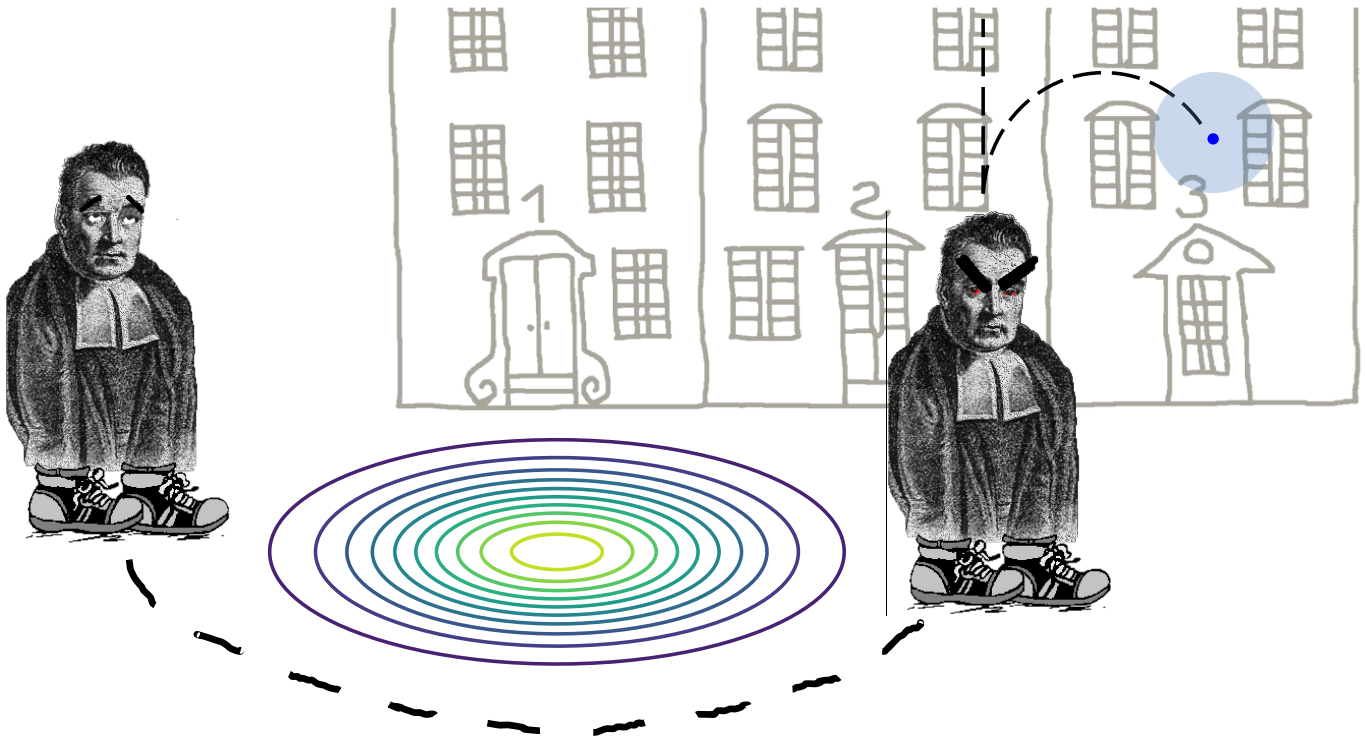

1
2
3

By comparison, prediction in "frequentist statistics" involves finding an optimum point estimate of the latent variables, e.g. by ML or MAP estimation, and then plugging this estimate into the likelihood distribution.

This has the disadvantage that it does not account for any uncertainty in the value of the parameter, and hence will underestimate the variance of the predictive distribution.

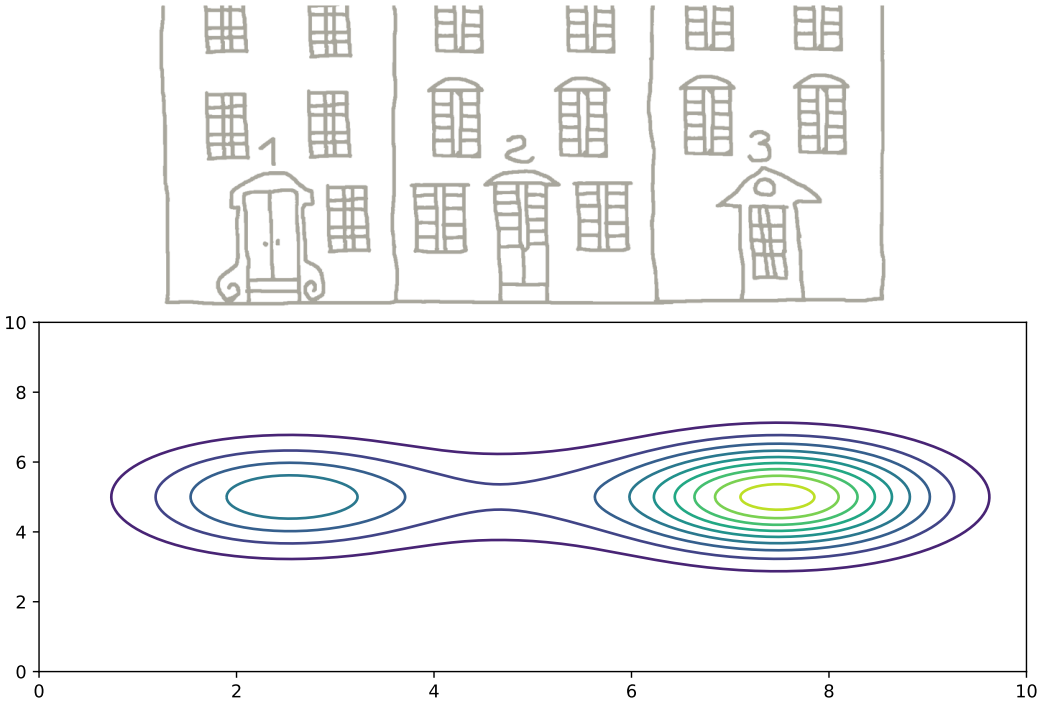
$$\hat{z}^{\text{MAP}} = \underset{i \in \{1,2,3\}}{\arg\max}\, p(z = i | \mathbf{x}_1, ..., \mathbf{x}_N) = 1$$

$$p(\mathbf{x}_{\text{new}} | z = \hat{z}^{\text{MAP}}) = \text{N}\left(\mathbf{x}_{\text{new}}; \boldsymbol{\mu}_1, \sigma^2 \mathbf{I}\right)$$


1
2
3

We can also compute the predictive prior, which tells us what we would predict given no observations. This is useful to check if the prior distribution does capture our prior beliefs.

Predictive prior

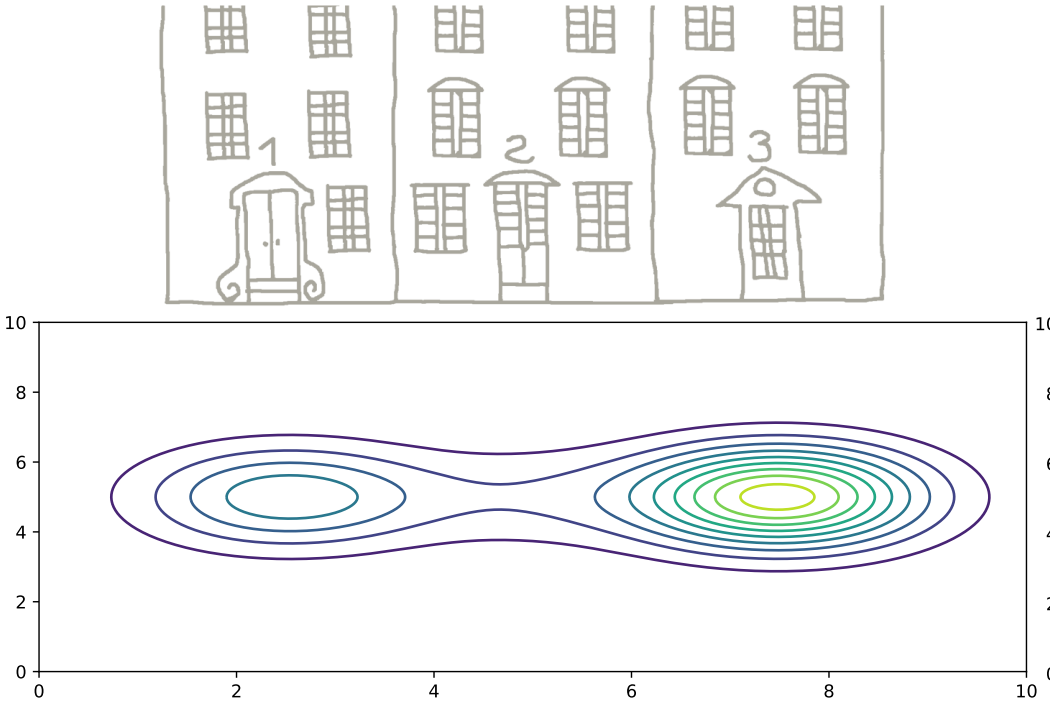


$$\mathrm{E}_{p(z=i)}\left[\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu}_i,\sigma^2\mathbf{I}\right)\right]$$

1st house: students; 2nd house: grandma; 3rd house: kids

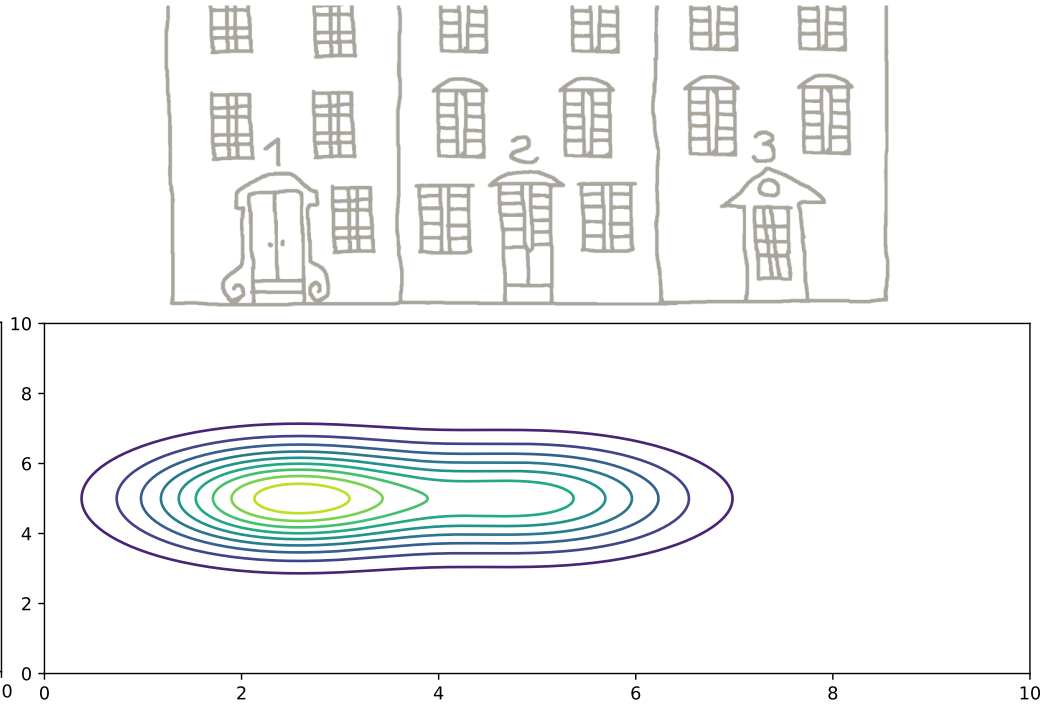
We can also compute the predictive prior, which tells us what we would predict given no observations. This is useful to check if the prior distribution does capture our prior beliefs.

Predictive prior



$$\mathrm{E}_{p(z=i)}\left[\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu}_i,\sigma^2\mathbf{I}\right)\right]$$

Predictive posterior



$$\mathrm{E}_{p(z=i|\mathbf{x}_1,\ldots,\mathbf{x}_N)}\left[\mathrm{N}\left(\mathbf{x};\boldsymbol{\mu}_i,\sigma^2\mathbf{I}\right)\right]$$

1st house: students; 2nd house: grandma; 3rd house: kids

Bayesian cooking recipe

Modeling

- What are the observations $\mathbf{x}$ (assumed continuous in R^D)?
- What are the latent (i.e. hidden, unobserved) variables $\mathbf{z}$ (assumed continuous in R^L)?
- Define the joint distribution:

$$p(\mathbf{x}, \mathbf{z}; \theta) = p(\mathbf{x}|\mathbf{z}; \theta_x)p(\mathbf{z}; \theta_z),$$

▷ $p(\mathbf{z}; \theta_z)$ is the prior, it encodes the prior bielief and uncertainty about the latent variables of interest (e.g. grandma jane is soooo kind, students become crazy when they party, kids like to play tricks).

▷ $p(\mathbf{x}|\mathbf{z}; \theta_x)$ is the likelihood, it defines how the observations are generated from the latent variables of interest (e.g. stones are launched independently from each other given the guilty house, and their position follows a Gaussian distribution).

▷ $\theta = \{\theta_x, \theta_z\}$ are the model (hyper-)parameters, which are deterministic.

Note that the likelihood is not a probability distribution over $\mathbf{z}$, its integral with respect to $\mathbf{z}$ does not (necessarily) equal one. The likelihood is a function of $\mathbf{z}$.

Inference

Inference consists in computing the posterior distribution of the latent variables $p(\mathbf{z}|\mathbf{x};\theta)$, i.e. the distribution of the latent random variables given the observations.

This distribution summarizes our knowledge on $\mathbf{z}$ once we have observed $\mathbf{x}$.

Using Bayes' theorem, the posterior distribution decomposes as:

$$p(\mathbf{z}|\mathbf{x};\theta) = \frac{p(\mathbf{x}|\mathbf{z};\theta_x)p(\mathbf{z};\theta_z)}{p(\mathbf{x};\theta)} = \frac{p(\mathbf{x}|\mathbf{z};\theta_x)p(\mathbf{z};\theta_z)}{\int p(\mathbf{x}|\mathbf{z};\theta_x)p(\mathbf{z};\theta_z)d\mathbf{z}},$$

where

$$p(\mathbf{x};\theta) = \int p(\mathbf{x},\mathbf{z};\theta_x)d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z};\theta_x)p(\mathbf{z};\theta_z)d\mathbf{z}$$

is called the marginal likelihood, or evidence.

In some cases, this integral cannot be computed analytically, in which case we need to resort to approximate inference techniques.

Decision

- The process of inference will often require us to use the posterior to answer various questions.
- $p(\mathbf{z}|\mathbf{x};\theta)$ contains all the information about $\mathbf{z}$ we care about, but it does not directly provide an "estimate" of the $\mathbf{z}$.

  From the posterior $p(\mathbf{z}|\mathbf{x};\theta)$, we need to choose a single value $\hat{\mathbf{z}}$ to serve as a point estimate of $\mathbf{z}$.

  To a Bayesian, this is a decision, and in different contexts we might want to select different point estimates.

- To take the decision, we need to introduce a loss function $l(\hat{\mathbf{z}}, \mathbf{z})$ which tells us "how bad" would $\hat{\mathbf{z}}$ be if the "true value" of the latent variable was $\mathbf{z}$.
- The decision is then taken by minimizing the posterior expected loss:

$$\mathrm{L}(\hat{\mathbf{z}}) = \mathrm{E}_{p(\mathbf{z}|\mathbf{x};\theta)}[l(\hat{\mathbf{z}}, \mathbf{z})].$$

- For example, let's consider the mean squared error (MSE) loss $l(\hat{\mathbf{z}}, \mathbf{z}) = (\hat{\mathbf{z}} - \mathbf{z})^2$.

  Minimizing the posterior expected loss with respect to $\hat{\mathbf{z}}$ gives the posterior mean:

$$\hat{\mathbf{z}}_{MSE} = \arg\min_{\hat{\mathbf{z}}} \mathrm{E}_{p(\mathbf{z}|\mathbf{x};\theta)}[(\hat{\mathbf{z}} - \mathbf{z})^2] = \mathrm{E}_{p(\mathbf{z}|\mathbf{x};\theta)}[\mathbf{z}].$$

  With different losses, we could obtain the MAP (also called the posterior mode), or the median.

Prediction

Predictive prior: "Averaging" the likelihood over the prior.

$$
\begin{aligned}
p(\mathbf{x}_{\text{new}}; \theta) &= \int p(\mathbf{x}_{\text{new}}, \mathbf{z}; \theta) d\mathbf{z} \\
&= \int p(\mathbf{x}_{\text{new}} | \mathbf{z}; \theta_x) p(\mathbf{z}; \theta_z) d\mathbf{z} \\
&= \mathrm{E}_{p(\mathbf{z}; \theta_z)} \left[p(\mathbf{x}_{\text{new}} | \mathbf{z}; \theta_x) \right].
\end{aligned}
$$

Predictive posterior: "Averaging" the likelihood over the posterior.

$$
\begin{aligned}
p(\mathbf{x}_{\text{new}} | \mathbf{x}; \theta) &= \int p(\mathbf{x}_{\text{new}}, \mathbf{z} | \mathbf{x}; \theta) d\mathbf{z} \\
&= \int p(\mathbf{x}_{\text{new}} | \mathbf{z}; \theta_x) p(\mathbf{z} | \mathbf{x}; \theta) d\mathbf{z} \\
&= \mathrm{E}_{p(\mathbf{z} | \mathbf{x}; \theta)} \left[p(\mathbf{x}_{\text{new}} | \mathbf{z}; \theta_x) \right].
\end{aligned}
$$

Bayesian inference for the Gaussian

Modeling

- We observe N independent and identically distributed (i.i.d) Gaussian variables:

$$\mathbf{x} = \{x_1, x_2, ..., x_N\}.$$

  Actually we observe N i.i.d realizations of one single Gaussian random variable.

- We suppose that the variance σ^2 is deterministic and known while the mean μ is treated as a latent variable.

- The modeling step consists in defining the joint distribution of the observed and latent variables, which factorizes as the product of the likelihood and the prior.

In the following, to ease the notations we omit the deterministic parameters of the distributions.

Likelihood

$$
\begin{aligned}
p(\mathbf{x}|\mu) &= p(x_1, x_2, ..., x_N|\mu) \\
&= \prod_{i=1}^{N} p(x_i|\mu) \\
&= \prod_{i=1}^{N} \mathrm{N}\left(x_i; \mu, \sigma^2\right) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i-\mu)^2\right]
\end{aligned}
$$

We recall that the ML estimate of the mean is given by $\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$ (proof let as an exercise).

Prior

We see that the likelihood function takes the form of the exponential of a quadratic form in μ.

Remember that the posterior is proportional to the product of the likelihood and the prior. Therefore, if we choose a Gaussian prior, the posterior will be a product of two exponentials of quadratic functions of μ and hence will be Gaussian too, making our Bayesian life easy:

$$p(\mu) = \mathrm{N}\left(\mu; \mu_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right],$$

where μ_0 and σ_0^2 are referred to as hyper-parameters.

Inference

Exercise

Applying Bayes' theorem, show that

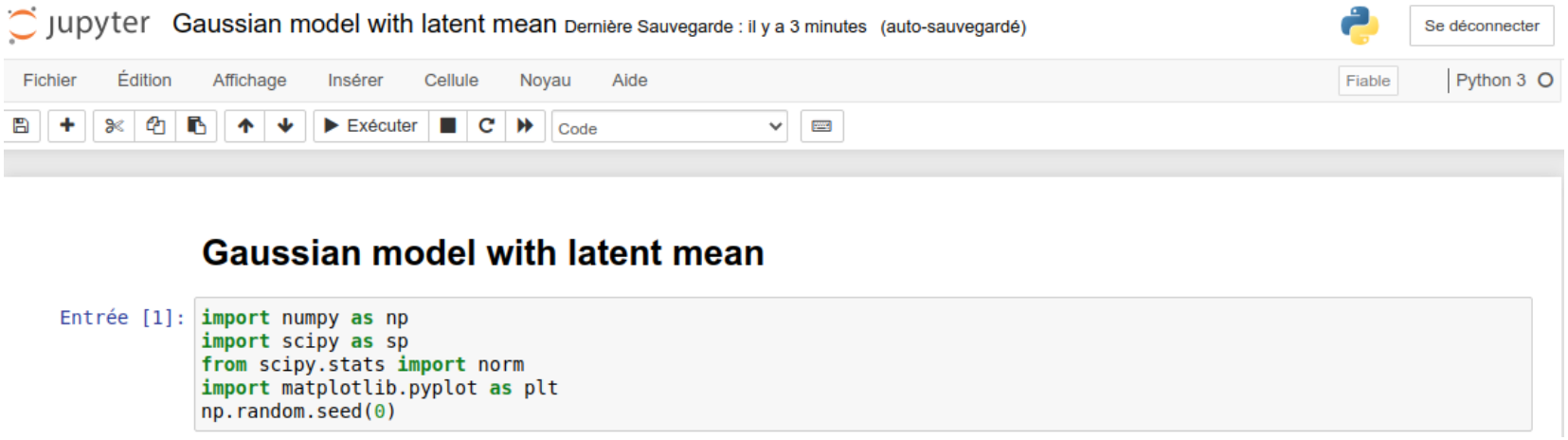
$$p(\mu|\mathbf{x}) = \mathrm{N}\left(\mu; \mu_\star, \sigma_\star^2\right),$$

where

- $\mu_\star = \dfrac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \dfrac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}$
- $\dfrac{1}{\sigma_\star^2} = \dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2}$

The inverse of the variance is called the precision.

Open the notebook "Gaussian model with latent mean.ipynb".



- The mean of the posterior distribution is a compromise between the prior mean and the maximum likelihood estimate.
- If the number of observations $N = 0$, the posterior mean and variance reduce to the prior mean and variance.
- For $N \to +\infty$, the posterior mean is given by the maximum likelihood solution. The prior has no effect in the "big data" regime.
- The precision of the posterior is given by the precision of the prior plus one contribution of the data precision from each of the observed data points. In other words, the precision linearly grows with the number of observed data points. The more data we have, the higher is the precision, the lower is the variance, the more certain we are about the MAP estimate.
- For $N \to +\infty$, the posterior variance goes to zero, and the posterior distribution becomes infinitely peaked around the maximum likelihood solution. ML point estimation is recovered from the Bayesian formalism in the limit of an infinite number of observations.
- For finite N, if we take the limit $\sigma_0^2 \to +\infty$ then the posterior mean reduces to the ML estimate and the posterior variance is given by σ^2/N. The prior is "not informative" if it is "too flat".

Homework

- Consider that mean is now a deterministic and known parameter while the variance is a latent random variable following an inverse-gamma prior distribution:

$$p(\sigma^2) = \mathrm{IG}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right),$$

  where $\Gamma(\cdot)$ is the Gamma function.

- The likelihood is still Gaussian: $p(\mathbf{x}|\sigma^2) = \prod_{i=1}^{N} \mathrm{N}\left(x_i; \mu, \sigma^2\right)$.

- Show that the posterior is given by:

$$p(\sigma^2|\mathbf{x}) = \mathrm{IG}(\sigma^2; \alpha_\star, \beta_\star),$$

  where $\alpha_\star = \alpha + \frac{N}{2}$ and $\beta_\star = \beta + \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2$.

The priors

- The inference, and therefore the predictions, decisions and actions which are based on the inference (posterior computation) all depend on the prior.
- The prior summarizes the information you have about the latent variables of interest, as well as the uncertainty related to this information.
- The prior is a key ingredient to Bayesian inference, but it is not sufficient, you also need data.
- The most often criticized aspect of the Bayesian approach to statistical inference is the requirement to choose a prior distribution, and especially the subjectivity of this prior selection procedure. But the whole modeling procedure is in any case inherently subjective, e.g. the likelihood modeling.

How to convert prior information into prior distributions?

Conjugate priors

Definition: A familiy of probability distributions is conjugate for a likelihood function if for every prior in this family the posterior also belongs to this familiy. In other words, if the posterior distribution is in the same family as the prior distribution, the prior is conjugate for the likelihood function.

- The "structure" of the prior is propagated to the posterior. Computing the posterior consists in updating the prior parameters using the observations.
- Using conjugate priors is simple and makes inference tractable (the marginal likelihood and therefore the posterior can be computed in closed form), but it is also constraining.
- For example, the Gaussian distribution is a conjugate prior for the Gaussian likelihood (with known variance); choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian.

Table of conjugate priors for various likelihood functions.

Non-informative priors

- A non-informative or uninformative prior is a prior distribution which is designed to have a weak influence the posterior distribution.
- It is useful when we do not have any clear prior beliefs about the latent variables of interest, or we do not want these prior beliefs to influence the inference process.
- A non-informative prior typically yields results which are not too different from conventional statistical analysis, as the likelihood function often yields more information than the non-informative prior.
- Remember the Bayesian Gaussian model with a high variance for the prior on the mean (cf. jupyter notebook).

Uniform prior

The simplest and oldest rule (suggested by the pioneers of the Bayesian inference, Bayes and Laplace) for determining a non-informative prior is the principle of indifference, which assigns equal probabilities to all possibilities.

This rule leads a uniform prior:

- Discrete case: $p(z) = 1/K$ for $z \in \{z_1, ..., z_K\}$.
- Continuous case: $p(z) \propto 1$ (constant pdf).

The uniform prior does not influence the posterior:

$$p(z|x) \propto p(x|z)p(z) \propto p(x|z).$$

- The uniform prior is not invariant under reparametrization:

  If we have no information about z, we also have no information about $y = g(z) = 1/z$. But a uniform prior over z does not correspond to a uniform prior over $1/z$.

  By the change of variable formula:

$$p_Y(y) = p_Z(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \propto y^{-2} \neq \textit{constant}$$

- The uniform prior is improper:

  If the random variable z is real-valued, the uniform prior $p(z) \propto 1$ does not integrate to 1, we say that it is improper.

  "Improperness" is not always a serious problem since improper priors can lead to proper posteriors.

Jeffreys' prior

In the univariate (i.e. 1-dimensional) case, it is defined by:

$$p(z) \propto \sqrt{I(z)},$$

where $I(z)$ is the Fisher information:

$$I(z) = \mathrm{E}_{p(x|z)}\left[\left(\frac{\partial \ln p(x|z)}{\partial z}\right)^2\right] = -\mathrm{E}_{p(x|z)}\left[\frac{\partial^2 \ln p(x|z)}{\partial^2 z}\right],$$

with $p(x|z)$ is the likelihood.

In the multivariate case, the prior is proportional to the square root of the determinant of the Fisher information matrix.

Jeffreys' prior is invariant under reparametrization:

$$
\begin{aligned}
p_Y(y) &= p_Z(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\
&\propto \sqrt{I(g^{-1}(y)) \left(\frac{d}{dy} g^{-1}(y)\right)^2} \\
&= \sqrt{\mathrm{E}_{p(x|y)} \left[\left(\frac{\partial \ln p(x|y)}{\partial g^{-1}(y)}\right)^2\right] \left(\frac{dg^{-1}(y)}{dy}\right)^2} \\
&= \sqrt{\mathrm{E}_{p(x|y)} \left[\left(\frac{\partial \ln p(x|y)}{\partial g^{-1}(y)} \frac{dg^{-1}(y)}{dy}\right)^2\right]} \\
&= \sqrt{\mathrm{E}_{p(x|y)} \left[\left(\frac{\partial \ln p(x|y)}{dy}\right)^2\right]} \\
&= \sqrt{I(y)}.
\end{aligned}
$$

Hierarchical prior

- Considering a conjugate prior $p(z;\theta_z)$ may be too restrictive.
- Instead of treating θ_z as a deterministic (hyper)parameters, we could consider it as another latent random variable, equipped with a prior $p(\theta_z;\lambda)$.
- The resulting prior over z is hierarchical, i.e. it is expressed as a marginal distribution:

$$p(z;\gamma) = \int p(z,\theta_z;\gamma)d\theta_z = \int p(z|\theta_z)p(\theta_z;\gamma)d\theta$$

- This prior is not conjugate anymore and it is "more expressive".

What we treat as a (hyper)parameter or as a random variable is quite arbitrary and problem-dependent.

Student's t distribution example

$$\begin{cases} p(x|v) & = \mathrm{N}\left(x;\mu,v\right) \\ p(v) & = \mathrm{IG}\left(v;\frac{\alpha}{2},\frac{\alpha}{2}\lambda^2\right) \end{cases} \quad \Leftrightarrow \quad p(x) = \int_0^{+\infty} p(x|v)p(v)dv = \mathrm{T}_{\alpha}(x;\mu,\lambda)$$

where

- $\mathrm{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$ is the Gaussian distribution .
- $\mathrm{IG}(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\beta/x\right)$ is the inverse-gamma distribution .
- $\mathrm{T}_{\alpha}(x;\mu,\lambda) = \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})\sqrt{\pi\alpha}\lambda} \left(1+\frac{1}{\alpha}\frac{(x-\mu)^2}{\lambda^2}\right)^{-\frac{\alpha+1}{2}}$ is the Student's t distribution.

Student's t distribution example

$$\begin{cases} p(x|v) & = \mathrm{N}\left(x;\mu,v\right) \\ p(v) & = \mathrm{IG}\left(v;\frac{\alpha}{2},\frac{\alpha}{2}\lambda^2\right) \end{cases} \quad \Leftrightarrow \quad p(x) = \int_0^{+\infty} p(x|v)p(v)dv = \mathrm{T}_{\alpha}(x;\mu,\lambda)$$

where

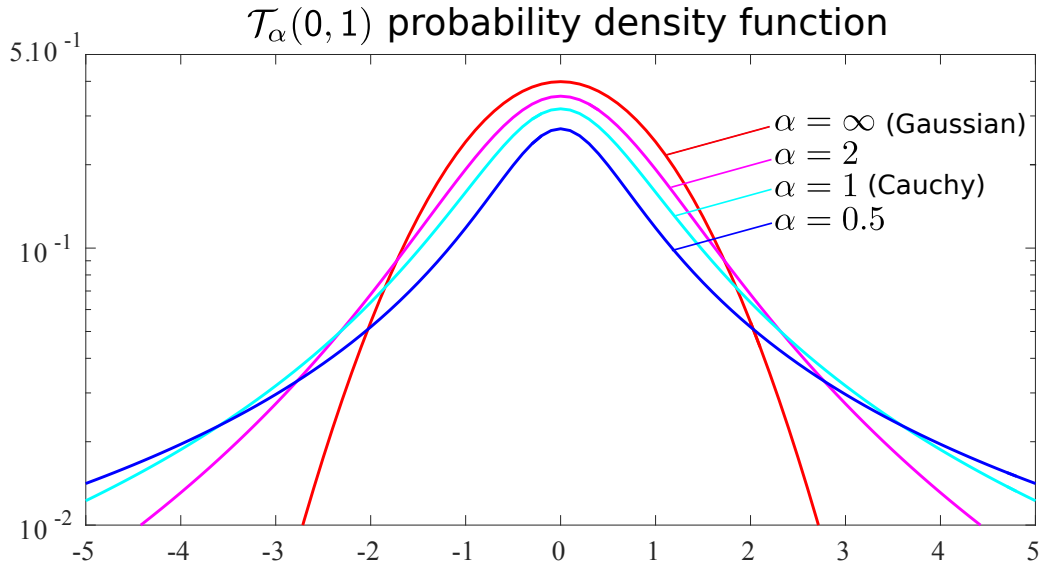
- $\mathrm{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$ is the Gaussian distribution .
- $\mathrm{IG}(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\beta/x\right)$ is the inverse-gamma distribution .
- $\mathrm{T}_{\alpha}(x;\mu,\lambda) = \frac{\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{\alpha}{2})\sqrt{\pi\alpha}\lambda} \left(1+\frac{1}{\alpha}\frac{(x-\mu)^2}{\lambda^2}\right)^{-\frac{\alpha+1}{2}}$ is the Student's t distribution.

Hints for the proof (exercise):

- 1st change of variable: $u = v\left(\frac{(x-u)^2}{2} + \frac{\alpha}{2}\lambda^2\right)$
- 2nd change of variable: $t = 1/u$
- Recognize the Gamma function $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}\,dt$ for $z = (\alpha+1)/2$.

Student's t distribution example

$$\begin{cases} p(x|v) &= \mathrm{N}\ (x; \mu, v) \\ p(v) &= \mathrm{IG}\ (v; \frac{\alpha}{2}, \frac{\alpha}{2}\lambda^2) \end{cases} \quad \Leftrightarrow \quad p(x) = \int_0^{+\infty} p(x|v)p(v)dv = \mathrm{T}_{\alpha}(x; \mu, \lambda)$$



It is a heavy-tailed distribution. It is more flexible than the Gaussian in the sense that it allows x to take values that are "far from the mode".

Natural image prior example

$$\mathbf{x} = \mathbf{H}\mathbf{z} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}).$$

- $\mathbf{x}$ is a noisy and/or incomplete image (e.g. motion blur, additive noise, missing pixels);
- $\mathbf{z}$ is the latent clean image;
- $\mathbf{H}$ encodes the linear transform (e.g. convolution) from the clean image to the noisy one;
- $\boldsymbol{\epsilon}$ is an additive Gaussian noise.

Image reconstruction (inpainting, denoising, deblurring) consists in computing $p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$.

$$\mathbf{x} = \mathbf{HGz} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right).$$

- $\mathbf{x}$ is a noisy and/or incomplete image (e.g. motion blur, additive noise, missing pixels);
- $\mathbf{z}$ is the latent horizonal gradient of the clean image;
- $\mathbf{G}^{-1}$ is the linear operator which computes the horizontal gradient of an image;
- $\mathbf{H}$ encodes the linear transform (e.g. convolution) from the clean image to the noisy one;
- $\boldsymbol{\epsilon}$ is an additive Gaussian noise.

Image reconstruction (inpainting, denoising, deblurring) consists in computing $p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ and somehow reconstruct the image from its gradient.

The likelihood is given by:

$$p(\mathbf{x}|\mathbf{z}) = \mathrm{N}\left(\mathbf{x}; \mathbf{HGz}, \sigma_\epsilon^2 \mathbf{I}\right).$$

To fully define the model, we also need a prior $p(\mathbf{z})$ for the gradient of the image.

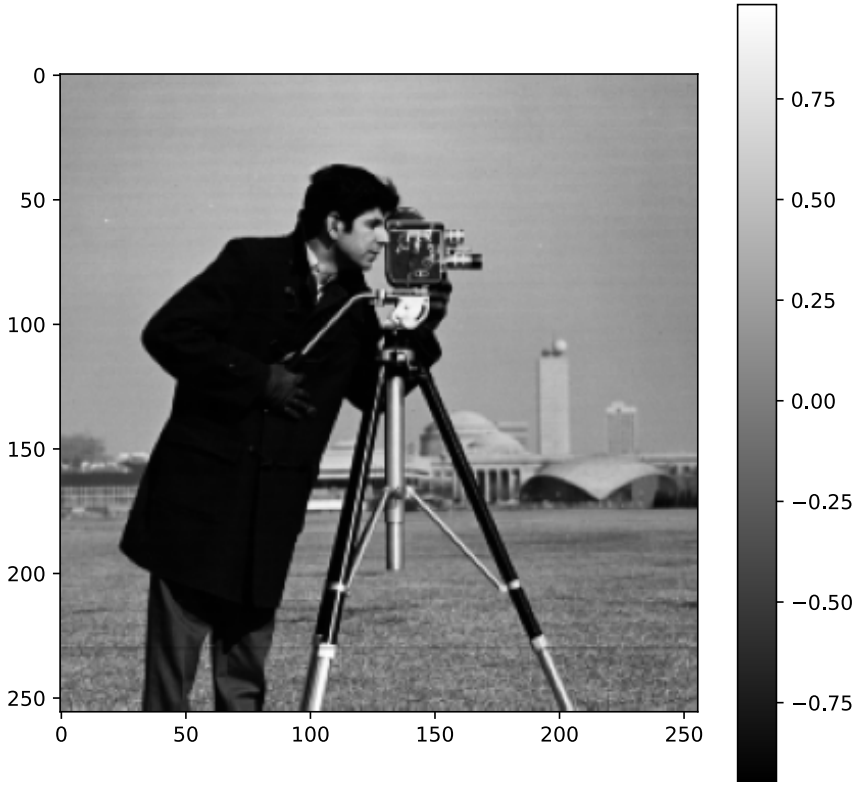
The conjugate prior for this likelihood function is Gaussian (assuming zero mean):

$$p(\mathbf{z}) = \prod_i \mathrm{N}\left(z_i; 0, \sigma_z^2\right).$$

Let's check if the distribution of the gradient of an image is indeed Gaussian.

We first load an image.

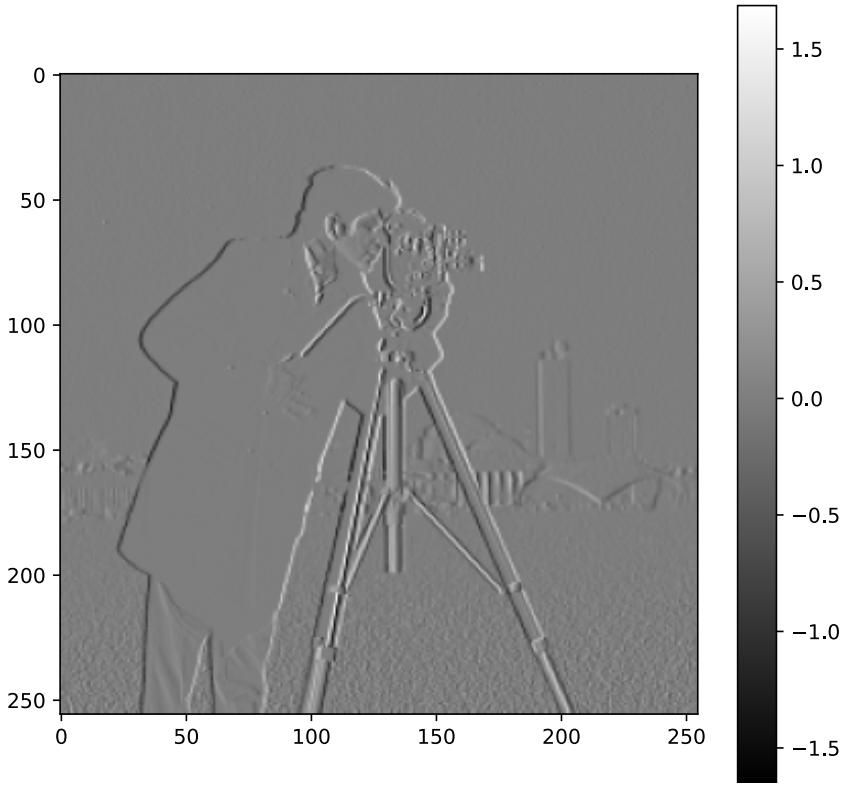
```
from PIL import Image
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t
from scipy.stats import norm*

im = np.array(Image.open('cameraman.png'))/255*2 - 1

plt.figure(figsize=(7,7))
plt.imshow(im, cmap='gray')
```



Then we compute the horizontal gradient.

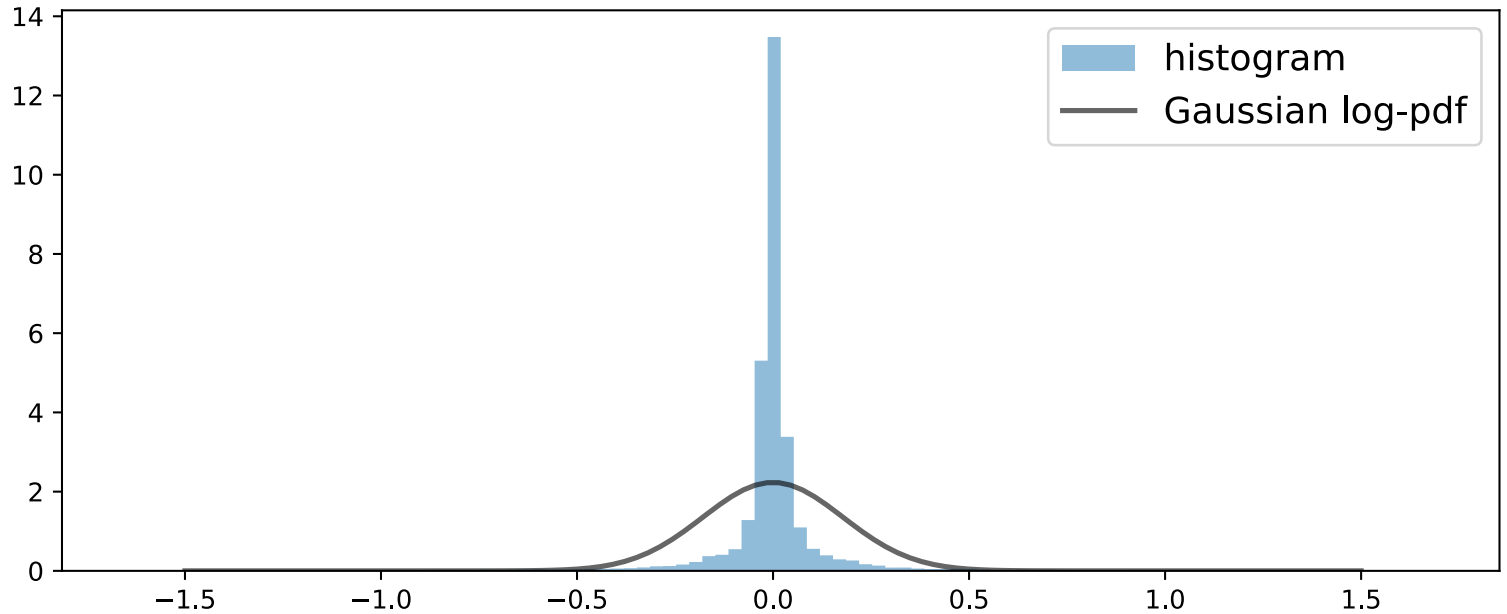
```
grad_x = np.diff(im)
grad_x_vec = grad_x.flatten()

plt.figure(figsize=(7,7))
plt.imshow(grad_x, cmap='gray')
plt.colorbar()
```

Can you guess what does the histogram of this image (once vectorized) look like?

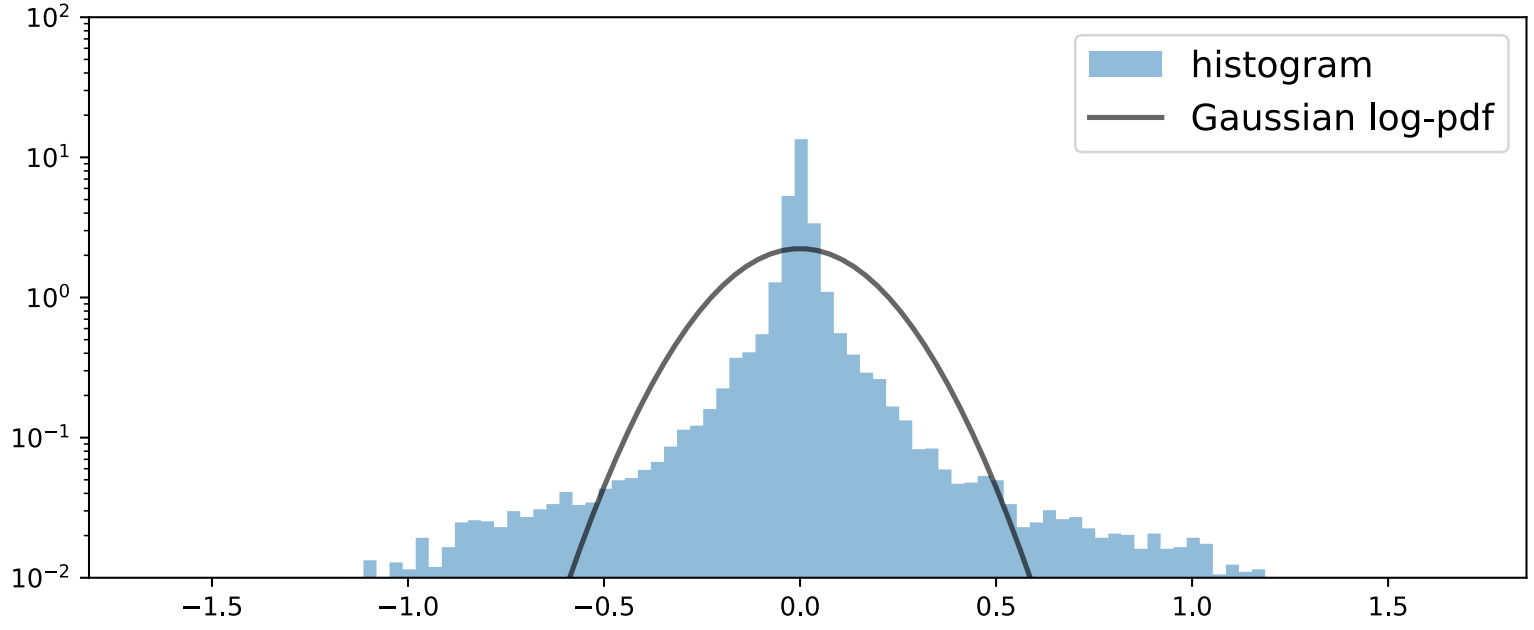


We fit a Gaussian distribution using maximum likelihood, and we compare the estimated pdf with the histogram of the gradient.

```
(mean, std) = norm.fit(grad_x_vec)
```



Same plot with a log scale on the y-axis.

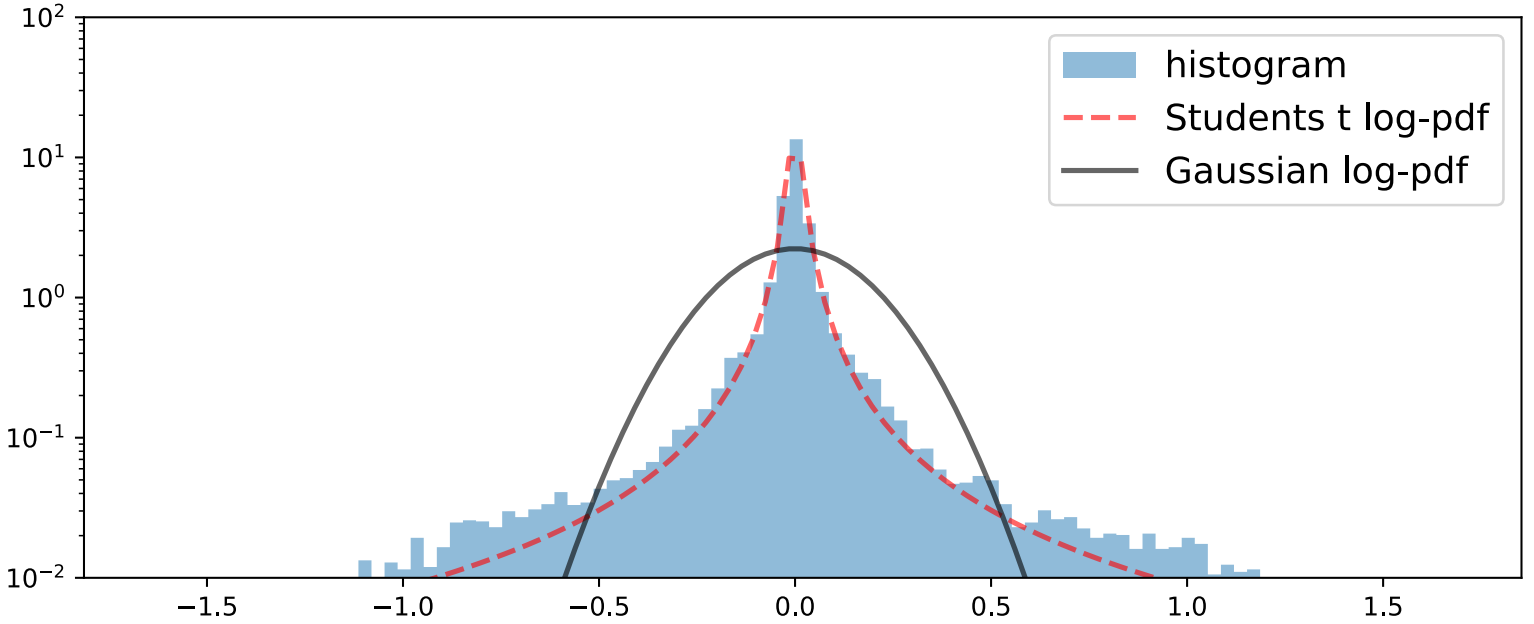


The distribution of the gradient of the image is clearly not Gaussian.

Choosing a Gaussian conjugate prior is too restrictive, and actually not a good idea here, so let's get hierarchical!

We fit a Student's t distribution on the gradient of the image.

```
(shape, loc, scale) = t.fit(grad_x_vec)
```



Much better!

Our hierarchical prior is $p(\mathbf{z}|\mathbf{v}) = \prod_i p(z_i|v_i)p(v_i)$, where

$$\begin{cases} p(z_i|v_i) & = \mathrm{N}\left(z_i; 0, v_i\right) \\ p(v_i) & = \mathrm{IG}\left(v_i; \frac{\alpha}{2}, \frac{\alpha}{2}\lambda^2\right) \end{cases},$$

which is equivalent to $p(\mathbf{z}) = \prod_i p(z_i)$ with

$$p(z_i) = \int_0^{+\infty} p(z_i|v_i)p(v_i)dv_i = \mathrm{T}_{\alpha}(z_i; 0, \lambda).$$

Variational Bayesian Sparse Kernel-Based Blind Image Deconvolution With Student's-t Priors

Dimitris G. Tzikas, Aristidis C. Likas, *Senior Member, IEEE*, and Nikolaos P. Galatsanos, *Senior Member, IEEE*

***Abstract*—In this paper, we present a new Bayesian model for the blind image deconvolution (BID) problem. The main novelty of this model is the use of a sparse kernel-based model for the point spread function (PSF) that allows estimation of both PSF shape and support. In the herein proposed approach, a robust model of the BID errors and an image prior that preserves edges of the reconstructed image are also used. Sparseness, robustness, and preservation of edges are achieved by using priors that are based on the Student's-t probability density function (PDF). This pdf, in addition to having heavy tails, is closely related to the Gaussian and, thus, yields tractable inference algorithms. The approximate variational inference methodology is used to solve the corresponding Bayesian model. Numerical experiments are presented that compare this BID methodology to previous ones using both simulated and real data.**

***Index Terms*—Bayesian approach, blind image deconvolution (BID), inverse problem, kernel model, sparse prior, student-t distribution.**

in image recovery problems and a book containing a review of the recent developments in mathematical tools for low level image processing problems can be found in [5] and [6], respectively. Methods based on anisotropic diffusion regularization have been also proposed [7]; however, they require the choice of the diffusion operator. There are also methods based on soft constraints [8], [9], which are very flexible; however, the form and the type of the used soft constraints is ad-hoc. Methods based on sparse image representations and quasi likelihood criteria have been also suggested [10].

Another way to apply constraints to the image and the PSF, is through the use of the Bayesian methodology. In this approach the unknown quantities are assumed to be random variables and suitable prior distributions are selected to impose the desired characteristics[11]–[16]. Unfortunately, since the BID data generation model is nonlinear, the posterior distribution of the un-

Example applications of Bayesian methods

Kaggle contest on Observing Dark World

From Observing Dark World:

> "There is more to the Universe than meets the eye. Out in the cosmos exists a form of matter that outnumbers the stuff we can see by almost 7 to 1, and we don't know what it is. What we do know is that it does not emit or absorb light, so we call it Dark Matter. Such a vast amount of aggregated matter does not go unnoticed. In fact we observe that this stuff aggregates and forms massive structures called Dark Matter Halos. Although dark, it warps and bends spacetime such that any light from a background galaxy which passes close to the Dark Matter will have its path altered and changed. This bending causes the galaxy to appear as an ellipse in the sky."

The contest required predictions about where dark matter was likely to be. The winner, Tim Salimans, used Bayesian inference to find the best locations for the halos (interestingly, the second-place winner also used Bayesian inference).

Tim Salimans' solution (source) :

- Construct a prior distribution for the halo positions $p(z)$, i.e. formulate our expectations about the halo positions before looking at the data.
- Construct a probabilistic model for the data (observed ellipticities of the galaxies) given the positions of the dark matter halos: $p(x|z)$.
- Use Bayes' rule to get the posterior distribution of the halo positions: $p(z|x)$, i.e. use the data to guess where the dark matter halos might be.
- Minimize the expected loss with respect to the posterior distribution over the predictions for the halo positions: $z^\star = \arg\min_{\hat{z}} \mathrm{E}_{p(z|x)}\left[\mathrm{L}(\hat{z}, z)\right]$, i.e. tune our predictions to be as good as possible for the given error metric.

The loss function in this problem is very complicated, it is about 160 lines of code, not something that can be written down in a single mathematical line.

Bayesian inference in gravitational-wave astronomy

Publications of the Astronomical Society of Australia (PASA)
doi: 10.1017/pas.2020.xxx.

An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models

v8 [astro-ph.IM] 12 Jun 2020

Eric Thrane[1,2,*] and Colm Talbot[1,2,†]
[1] Centre for Astrophysics, School of Physics and Astronomy, Monash University, VIC 3800, Australia
[2]OzGrav: The ARC Centre of Excellence for Gravitational-Wave Discovery, Clayton, VIC 3800, Australia

Abstract

This is an introduction to Bayesian inference with a focus on hierarchical models and hyper-parameters. We write primarily for an audience of Bayesian novices, but we hope to provide useful insights for seasoned veterans as well. Examples are drawn from gravitational-wave astronomy, though we endeavor for the presentation to be understandable to a broader audience. We begin with a review of the fundamentals: likelihoods, priors, and posteriors. Next, we discuss Bayesian evidence, Bayes factors, odds, and model selection. From there, we describe how posteriors are estimated using samplers such as Markov Chain Monte Carlo algorithms and nested sampling. Finally, we generalize the formalism to discuss hyper-parameters and hierarchical models. We include extensive appendices discussing the creation of credible intervals, Gaussian noise, explicit marginalization, posterior predictive distributions, and selection effects.

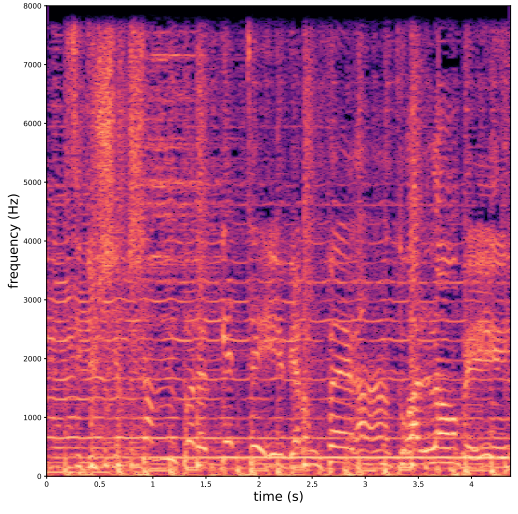
Keywords: gravitational waves – Bayesian inference – parameter estimation – model selection – hierarchical modeling

Bayesian audio-visual multi-speaker tracking

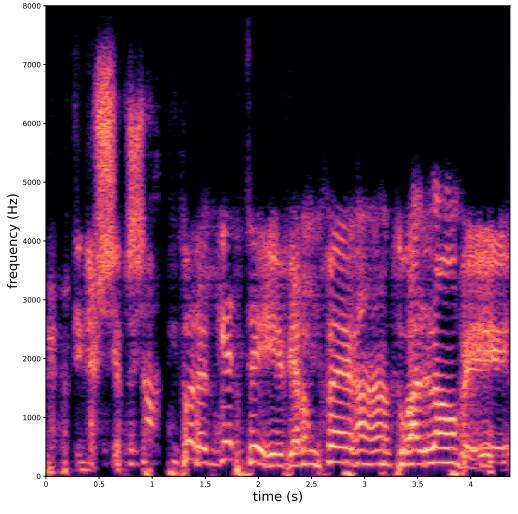
Y. Ban et al., Variational Bayesian Inference for Audio-Visual Tracking of Multiple Speakers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

Bayesian music source separation with a deep speech prior

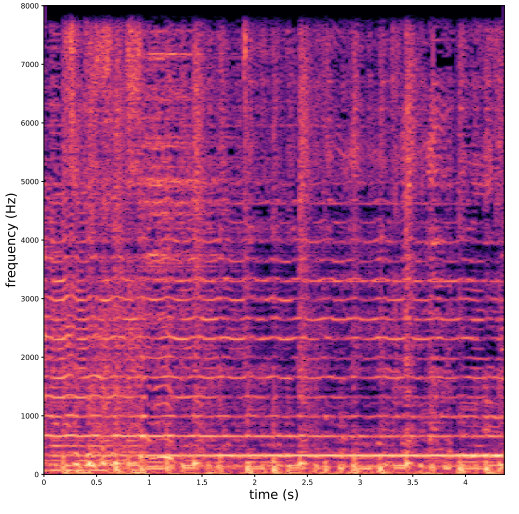
Mix



Vocals



Accompaniment



S. Leglaive et al., A recurrent variational autoencoder for speech enhancement, IEEE ICASSP 2020.