DTU

Model-based Machine Learning

PGM foundations II

Outline

PGMs in continuous domain

Generative processes

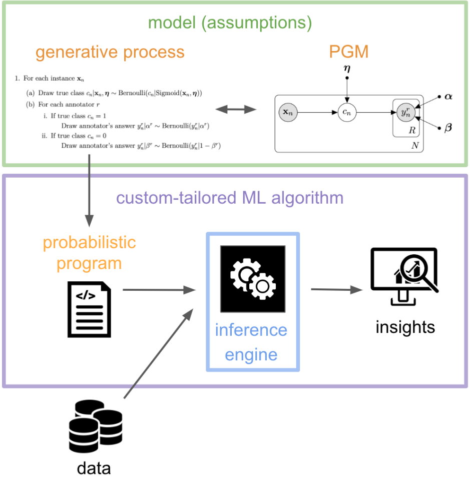
(Based on Michael Jordan, David Blei)

Road to MBML: where are we?

Foundations of PGMs and probabilistic programming
Probabilistic modelling in various contexts - your modelling toolbox
Bayesian inference (exact and approximate)

Week	Topic
1	Intro to the course + Prob. review
2	PGM foundations
3	PGM foundations II
4	Freq. vs Bayesian + Prob. Prog. + Mixture models
5	Regression models
6	Classification and Hierarchical models
7	Temporal models
8	Topic Models
9	Markov-chain Monte Carlo (MCMC)
10	Variational inference
11	Generative models
12	Gaussian processes
13	Project support

Model-based Machine Learning



Learning objectives

At the end of this lecture, you should be able to:

- Explain the concept of continuous random variable and its specification in a PGM
- Explain the concept of Bayesian inference and its relation to Bayes' theorem
- Explain the role of the likelihood, prior and model evidence in Bayes' theorem
- Explain the role of the prior, the importance of its form, and the concept of conjugate prior in inference
- Apply the generative process principles in the creation of a PGM and perform ancestral sampling with it

PGM in continuous domain

- Thus far, we've been using only discrete variables
- Conditional Probability Tables
- Extension to continuous domain is intuitive...

PGM in continuous domain

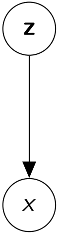
- Thus far, we've been using only discrete variables
- Conditional Probability Tables
- Extension to continuous domain is intuitive...
- But with it, some concepts become more relevant

PGM in continuous domain

- Thus far, we've been using only discrete variables
- Conditional Probability Tables
- Extension to continuous domain is intuitive...
- But with it, some concepts become more relevant
  - Prior
  - Conjugate prior

PGMs in continuous domain

- General form:

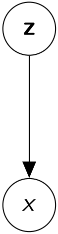


- We use functions f instead of tables to describe relationships between variables
- We typically assume that each random variable follows a well-known distribution (or combination of them) parameterized by a set $\boldsymbol{\theta}$. For example:

$$x \sim \text{Exponential}(\boldsymbol{\theta})$$

PGMs in continuous domain

- General form:



- We use functions f instead of tables to describe relationships between variables
- We typically assume that each random variable follows a well-known distribution (or combination of them) parameterized by a set $\boldsymbol{\theta}$. For example:

$$x \sim \text{Exponential}(\boldsymbol{\theta})$$

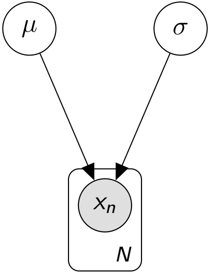
- The parameters of that distribution are a function of their "parent" variables $\mathbf{z}$

$$\boldsymbol{\theta} = f(\mathbf{z})$$

- f can be any function (albeit ideally differentiable - *why?*) of its inputs (identity, linear, polynomial, neural network, etc.)

PGMs in continuous domain

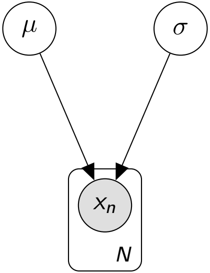
- A concrete well-known example: Gaussian distribution



- In this PGM, we assume to have observations x_n, that follow a Gaussian distribution
- It has two parameters (mean μ, variance σ^2)

PGMs in continuous domain

- A concrete well-known example: Gaussian distribution



- In this PGM, we assume to have observations x_n, that follow a Gaussian distribution
- It has two parameters (mean μ, variance σ^2)
- Fitting (parameter estimation):
  - It has a well-known likelihood function

$$L(\mu, \sigma) = \prod_{i}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

  - Corresponding log version

$$LL(\mu, \sigma) = -\frac{N}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$

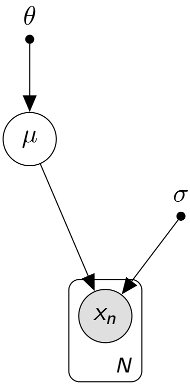
PGMs in continuous domain

- A Graphical Model allows for a full Bayesian treatment
  - We can assign *priors* to the parameters
  - We can use domain knowledge
  - Good to prevent overfitting

PGMs in continuous domain

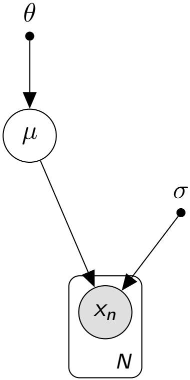
- A Graphical Model allows for a full Bayesian treatment
  - We can assign *priors* to the parameters
  - We can use domain knowledge
  - Good to prevent overfitting
  - What would be the form of those priors?

Gaussian distribution case



- To simplify, let's assume we know σ but not μ

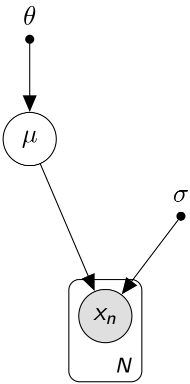
Gaussian distribution case



- To simplify, let's assume we know σ but not μ
- Can we pick *any* distribution, $p(\mu|\theta)$?
- Our joint distribution would become:

$$p(\mu, \mathbf{x}|\theta, \sigma) = p(\mu|\theta) \prod_{n=1}^{N} p(x_n|\mu, \sigma)$$

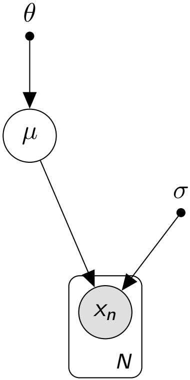
Gaussian distribution case



- To simplify, let's assume we know σ but not μ
- Can we pick *any* distribution, $p(\mu|\theta)$?
- **Common simplification** to unclutter notation:

$$p(\mu, \mathbf{x}|\theta, \sigma) = p(\mu|\theta) \prod_{n=1}^{N} p(x_n|\mu, \sigma)$$

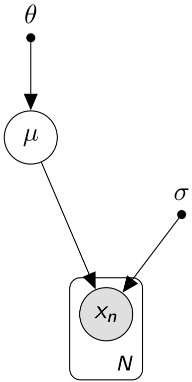
Gaussian distribution case



- To simplify, let's assume we know σ but not μ
- Can we pick *any* distribution, $p(\mu|\theta)$?
- **Common simplification** to unclutter notation:

$$p(\mu, \mathbf{x}) = p(\mu|\theta) \prod_{n=1}^{N} p(x_n|\mu, \sigma)$$

Gaussian distribution case

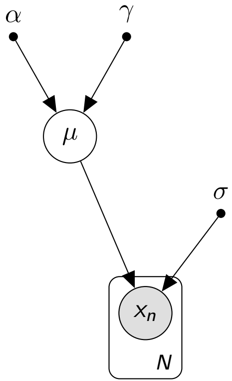


- To simplify, let's assume we know σ but not μ
- Can we pick *any* distribution, $p(\mu|\theta)$?
- Our joint distribution would become:

$$p(\mu, \mathbf{x}) = p(\mu|\theta) \prod_{n=1}^{N} p(x_n|\mu, \sigma)$$

- If $p(\mu|\theta)$ is normal, then $p(\mu, \mathbf{x})$ is normal too!

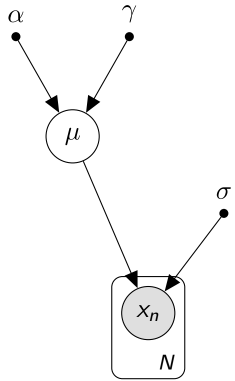
Gaussian distribution case



- If $p(\mu|\theta)$ is normal, then $p(\mu, \mathbf{x})$ is normal too!

$$p(\mu|\theta) = \mathcal{N}(\mu|\alpha, \gamma)$$

Gaussian distribution case



- If $p(\mu|\theta)$is normal, then $p(\mu, \mathbf{x})$ is normal too!

$$p(\mu|\theta) = \mathcal{N}(\mu|\alpha, \gamma)$$

- the log probability of our PGM would be:

$$\begin{aligned} LL(\mu, \alpha, \gamma, \sigma) = -\frac{N}{2}(\log(2\pi) + \log(\sigma)) \\ -\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 \\ -\frac{\log(2\pi)}{2} - \frac{\log(\gamma^2)}{2} - \frac{(\alpha - \mu)^2}{2\gamma^2} \end{aligned}$$

Playtime!

- Open notebook "03 - PGM fundamentals.ipynb"
- Do part 1 (est. duration=30 min)

Conjugate priors

- For many known distributions, there is a corresponding *conjugate prior*, P, that preserves its form under multiplication. I.e., if we have distribution L and its conjugate prior P_0, we should have

$$P_1 = L \times P_0,$$

  where P_1 has the same form as P_0
- For example, the Beta distribution is the conjugate prior of Bernoulli; and we've seen that the Normal is the conjugate for the mean of the Normal (when variance is known).
- If we have a known closed form for model, inference is generally more efficient!

Conjugate priors

- For many known distributions, there is a corresponding *conjugate prior*, P, that preserves its form under multiplication. I.e., if we have distribution L and its conjugate prior P_0, we should have

$$P_1 = L \times P_0,$$

  where P_1 has the same form as P_0
- For example, the Beta distribution is the conjugate prior of Bernoulli; and we've seen that the Normal is the conjugate for the mean of the Normal (when variance is known).
- If we have a known closed form for model, inference is generally more efficient!
- **This is great for online learning (why?)!**

Parameter inference: Beta–Bernoulli (coin bias)

A concrete example: Beta–Bernoulli (coin bias)

- **Model.** Unknown coin bias $\theta \in [0,1]$. Observations $x_1, \ldots, x_N$, $x_n \in \{0,1\}$

$$\text{Prior: } p(\theta) = \mathrm{Beta}(\alpha, \beta), \qquad \text{Likelihood: } p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \mathrm{Bernoulli}(x_n|\theta)$$

Parameter inference: Beta–Bernoulli (coin bias)

A concrete example: Beta–Bernoulli (coin bias)

- **Model.** Unknown coin bias $\theta \in [0, 1]$. Observations $x_1, \ldots, x_N$, $x_n \in \{0, 1\}$

$$\text{Prior: } p(\theta) = \mathrm{Beta}(\alpha, \beta), \qquad \text{Likelihood: } p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \mathrm{Bernoulli}(x_n|\theta)$$

- Let $s = \sum_{n=1}^{N} x_n$ be #heads, and $f = N - s$ be #tails. Then:

$$\textbf{Posterior: } p(\theta|x_{1:N}) = \mathrm{Beta}(\alpha + s,\ \beta + f)$$

Parameter inference: Beta–Bernoulli (coin bias)

A concrete example: Beta–Bernoulli (coin bias)

- **Model.** Unknown coin bias $\theta \in [0, 1]$. Observations $x_1, \ldots, x_N$, $x_n \in \{0, 1\}$

$$\text{Prior: } p(\theta) = \mathrm{Beta}(\alpha, \beta), \qquad \text{Likelihood: } p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \mathrm{Bernoulli}(x_n|\theta)$$

- Let $s = \sum_{n=1}^{N} x_n$ be #heads, and $f = N - s$ be #tails. Then:

$$\textbf{Posterior: } p(\theta|x_{1:N}) = \mathrm{Beta}(\alpha + s,\ \beta + f)$$

- **Interpretation (pseudo-counts):**

$$\alpha \text{ acts like "prior heads + 1"}, \quad \beta \text{ acts like "prior tails + 1"}$$

- So the data simply *adds* counts to the prior

Beta–Bernoulli is not magic: derive the posterior

- **Goal:** show that prior $\times$ likelihood has the form of a Beta again

Beta–Bernoulli is not magic: derive the posterior

- **Goal:** show that prior $\times$ likelihood has the form of a Beta again
- **Prior (Beta):**

$$p(\theta) = \mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in (0,1)$$

Beta–Bernoulli is not magic: derive the posterior

- **Goal:** show that prior $\times$ likelihood has the form of a Beta again
- **Prior (Beta):**

$$p(\theta) = \mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in (0, 1)$$

- **Data model (Bernoulli):** for $x_n \in \{0, 1\}$,

$$p(x_n \mid \theta) = \theta^{x_n}(1-\theta)^{1-x_n}$$

Beta–Bernoulli is not magic: derive the posterior

- **Goal:** show that prior $\times$ likelihood has the form of a Beta again
- **Prior (Beta):**

$$p(\theta) = \mathrm{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in (0,1)$$

- **Data model (Bernoulli):** for $x_n \in \{0,1\}$,

$$p(x_n \mid \theta) = \theta^{x_n}(1-\theta)^{1-x_n}$$

- **Bayes rule (up to proportionality):**

$$p(\theta \mid x_{1:N}) \propto p(x_{1:N} \mid \theta)\, p(\theta)$$

Derivation: multiply and collect exponents

- **Likelihood for i.i.d. Bernoulli flips:**

$$p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\sum_n x_n}(1-\theta)^{\sum_n (1-x_n)}$$

Derivation: multiply and collect exponents

- **Likelihood for i.i.d. Bernoulli flips:**

$$p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\sum_n x_n}(1-\theta)^{\sum_n (1-x_n)}$$

- Let $s = \sum_{n=1}^{N} x_n$ (#heads), and $f = \sum_{n=1}^{N}(1-x_n) = N - s$ (#tails)
- Then:

$$p(x_{1:N} \mid \theta) = \theta^{s}(1-\theta)^{f}$$

Derivation: multiply and collect exponents

- **Likelihood for i.i.d. Bernoulli flips:**

$$p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\sum_n x_n}(1-\theta)^{\sum_n (1-x_n)}$$

- Let $s = \sum_{n=1}^{N} x_n$ (#heads), and $f = \sum_{n=1}^{N}(1-x_n) = N - s$ (#tails)
- Then:

$$p(x_{1:N} \mid \theta) = \theta^s(1-\theta)^f$$

- **Posterior (unnormalized):**

$$\begin{aligned} p(\theta \mid x_{1:N}) &\propto \underbrace{\theta^s(1-\theta)^f}_{\text{likelihood}} \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \\ &= \theta^{(\alpha+s)-1}(1-\theta)^{(\beta+f)-1} \end{aligned}$$

- **Key step:** exponents add $\Rightarrow$ same functional family

Normalization: identify the Beta form

- We just showed

$$p(\theta \mid x_{1:N}) \propto \theta^{(\alpha+s)-1}(1-\theta)^{(\beta+f)-1}$$

- But a Beta density has the form

$$\mathrm{Beta}(\theta|a, b) = \frac{1}{B(a, b)}\,\theta^{a-1}(1-\theta)^{b-1}$$

Normalization: identify the Beta form

- We just showed

$$p(\theta \mid x_{1:N}) \propto \theta^{(\alpha+s)-1}(1-\theta)^{(\beta+f)-1}$$

- But a Beta density has the form

$$\mathrm{Beta}(\theta|a,b) = \frac{1}{B(a,b)}\,\theta^{a-1}(1-\theta)^{b-1}$$

- So we can *read off* the posterior parameters:

$$\boxed{\theta \mid x_{1:N} \sim \mathrm{Beta}(\alpha+s,\ \beta+f)}$$

Beta–Bernoulli: strength of the prior (same data, different priors)

Assume we observe $N = 10$ flips with $s = 8$ heads and $f = 2$ tails

Prior	(α, β)	Posterior $(\alpha + s, \beta + f)$	Posterior mean $\mathbb{E}[\theta \mid x]$
Weak-ish, centered	(2,2)	(10,4)	$\frac{10}{14} \approx 0.714$
Strong, centered	(20,20)	(28,22)	$\frac{28}{50} = 0.56$
Strong, skewed	(20,5)	(28,7)	$\frac{28}{35} = 0.80$

(recall that, if $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$)

Beta–Bernoulli: strength of the prior (same data, different priors)

Assume we observe $N = 10$ flips with $s = 8$ heads and $f = 2$ tails

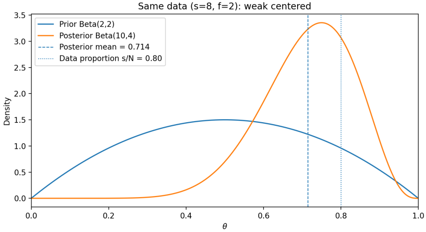
Prior	(α, β)	Posterior $(\alpha + s, \beta + f)$	Posterior mean $\mathbb{E}[\theta \mid x]$
Weak-ish, centered	(2,2)	(10,4)	$\frac{10}{14} \approx 0.714$
Strong, centered	(20,20)	(28,22)	$\frac{28}{50} = 0.56$
Strong, skewed	(20,5)	(28,7)	$\frac{28}{35} = 0.80$

(recall that, if $\theta \sim \mathrm{Beta}(\alpha, \beta)$, then $\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$)

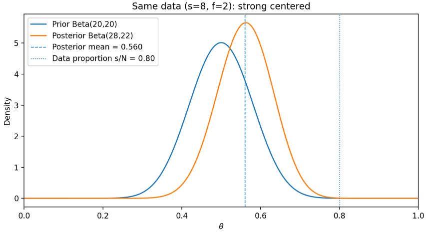
Discussion points:

- Which prior moves the least, and why?
- What does "prior strength" means?

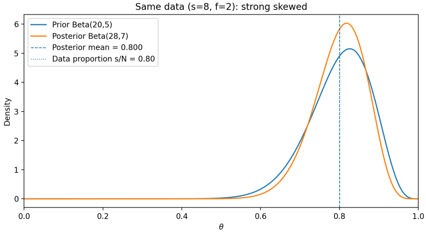
Same data, different priors: prior vs posterior shapes



Prior Beta$(2, 2) \rightarrow$ Posterior Beta$(10, 4)$



Prior Beta$(20, 20) \rightarrow$ Posterior Beta$(28, 22)$



Prior Beta$(20, 5) \rightarrow$ Posterior Beta$(28, 7)$

- Weak prior $\Rightarrow$ posterior tracks the data closely
- Strong prior $\Rightarrow$ posterior moves less (needs more data)
- Skewed strong prior $\Rightarrow$ posterior remains biased unless data is overwhelming

Conjugate priors

- We usually use a table

Discrete distributions [edit]

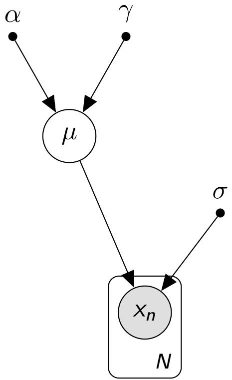
Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			$\alpha,\ \beta$[note 3]	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{\alpha_i{}'}{\sum_i \alpha_i{}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$

Figure: From Wikipedia

Some conjugate priors to remember…

Likelihood	**Prior**
Normal with known variance	Normal
Normal with known mean	Inverse Gamma
Multivariate normal, known mean	Inverse Wishart
Multivariate normal, unknown mean and variance	Normal-inverse-Wishart
Exponential	Gamma
Bernoulli	Beta
Mulitnomial	Dirichlet
Poisson	Gamma

Gaussian distribution case



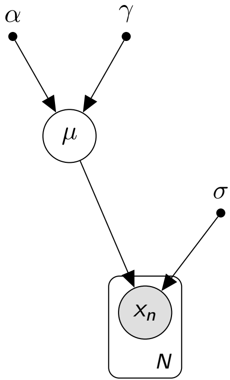
- For our Gaussian example, the posterior $p(\mu|\mathbf{x}) = \mathcal{N}(\tilde{\alpha}, \tilde{\gamma})$ will be directly

$$\tilde{\alpha} = \frac{1}{\gamma^{-2} + \frac{N}{\sigma^2}} \left(\frac{\alpha}{\gamma^2} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2} \right)$$

$$\tilde{\gamma} = \sqrt{\left(\gamma^{-2} + \frac{N}{\sigma^2} \right)^{-1}}$$

- We just followed the conjugate priors table
- Calculation in constant time, no need to optimize anything!
- We could use this as the next prior!...

Gaussian distribution case



- For our Gaussian example, the posterior $p(\mu|\mathbf{x}) = \mathcal{N}(\tilde{\alpha}, \tilde{\gamma})$ will be directly

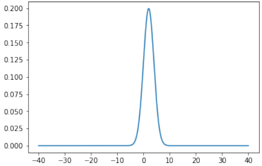
$$\tilde{\alpha} = \frac{1}{\gamma^{-2} + \frac{N}{\sigma^2}} \left(\frac{\alpha}{\gamma^2} + \frac{\sum_{i=1}^{N} x_i}{\sigma^2} \right)$$

$$\tilde{\gamma} = \sqrt{\left(\gamma^{-2} + \frac{N}{\sigma^2} \right)^{-1}}$$

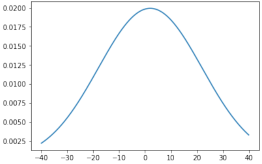
- We just followed the conjugate priors table
- Calculation in constant time, no need to optimize anything!
- We could use this as the next prior!...
- BUT if $p(\mu, \mathbf{x})$ is not a known distribution, we may have trouble deriving it (analytically)...

Last note on priors

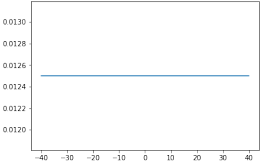
- Depending on what you know of the problem (or the constraints you want to impose...):



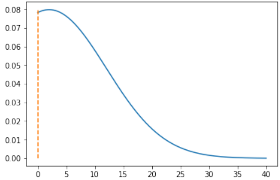
Proper informative prior



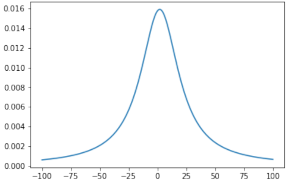
Proper weakly informative prior



Inproper uninformative prior



Proper bounded prior



Proper uninformative prior (notice scale and fat tails)

Generative processes

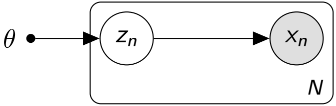
- By now, you understand that you can combine variables in multiple ways in your graphical model
- On the other hand, you may be overwhelmed about where to start doing your own
  - Small models, with few variables, are simple
  - What if you have a lot of variables, assumptions, domain knowledge?...

Generative processes

- By now, you understand that you can combine variables in multiple ways in your graphical model
- On the other hand, you may be overwhelmed about where to start doing your own
  - Small models, with few variables, are simple
  - What if you have a lot of variables, assumptions, domain knowledge?...
- You need to think from a generative perspective...

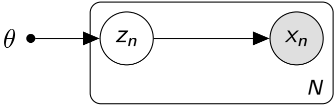
"Generative story" of data

- How is a data point generated?



"Generative story" of data

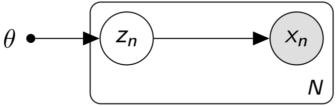
- How is a data point generated?



- Set parameter θ (fixed - not a random variable!)
- For $n = 1..N$, do
  1. Draw a random latent variable, $z_n \sim p(z_n|\theta)$

"Generative story" of data

- How is a data point generated?



- Set parameter θ (fixed - not a random variable!)
- For $n = 1..N$, do
  1. Draw a random latent variable, $z_n \sim p(z_n|\theta)$
  2. Given z_n, draw x_n such that $x_n \sim p(x_n|z_n)$

- In fact, this resembles a program structure!

A more complex example - Dwell time prediction

For a given bus stop, that serves a single line, can we predict the amount of time the next bus will be stopped there to load/unload passengers (the *dwell* time)?

- Our dataset contains $\{x_n = \{0, 1\}$-representing peak/non-peak hour, dt_n - dwell time$\}$.
- Notice that, sometimes, the bus does not stop at all!
- When it stops, we measure the duration as dt
- When it doesn't stop, $dt = 0$

Dwell time prediction

Given fixed parameters: σ_β, σ_ϵ and π

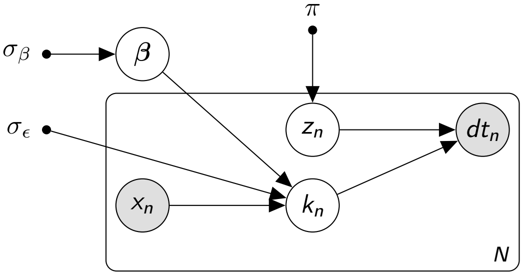
1. Draw a pair of parameters[1]: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta \mathbf{I})$
2. For $n = \{1, \ldots, N\}$
   1. Draw one value for k_n, such that $k_n \sim \mathcal{N}(\beta_0 + \beta_1 x_n, \sigma_\epsilon)$
   2. Draw $z_n \sim \text{Bernoulli}(\pi)$ - (you can think of this as flipping a biased coin)
   3. If $z_n = 1$, then bus has stopped
      - The bus has stopped, so set dwell-time $dt_n = k_n$
   4. Else:
      - The bus didn't stop, so set dwell-time $dt_n = 0$

[1] We need two values for β, one for the intercept, another for the peak/non-peak information.

Dwell time prediction

Given fixed parameters: σ_β, σ_ϵ and π

1. Draw a pair of parameters[1]: $\beta \sim \mathcal{N}(\mathbf{0}, \sigma_\beta \mathbf{I})$
2. For $n = \{1, \ldots, N\}$
   1. Draw one value for k_n, such that $k_n \sim \mathcal{N}(\beta_0 + \beta_1 x_n, \sigma_\epsilon)$
   2. Draw $z_n \sim \text{Bernoulli}(\pi)$ - (you can think of this as flipping a biased coin)
   3. If $z_n = 1$, then bus has stopped
      - The bus has stopped, so set dwell-time $dt_n = k_n$
   4. Else:
      - The bus didn't stop, so set dwell-time $dt_n = 0$



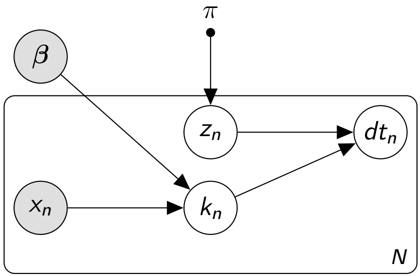
[1] We need two values for β, one for the intercept, another for the peak/non-peak information.

Dwell time prediction

- After you define your model, you need to estimate it. I.e. infer the following:
  - Distribution of β
  - Optimal values of σ_ϵ, σ_β, and π (we defined them as constants!)
- Of course, when you have them, you can make your predictions!
- Your model will look different:

Dwell time prediction

- After you define your model, you need to estimate it. I.e. infer the following:
  - Distribution of β
  - Optimal values of σ_ϵ, σ_β, and π (we defined them as constants!)
- Of course, when you have them, you can make your predictions!
- Your model will look different:



"Generative story" of data

- Set up the building blocks, as per available knowledge
- Easy to change data distributions inside the model
- Can be used to *actually* generate data!
  - Ancestral sampling
  - Do *prior predictive checks*!

Playtime!

- Open notebook "03 - PGM fundamentals.ipynb"
- Do part 2 (est. duration=30 min)

Readings

- **Main reading**: Chapter 8: “Graphical Models”, pages 359-366 and pages 372-379 of Chris Bishop’s book, “Pattern Recognition and Machine Learning” (PRML) URL: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf)
- “D-separation: How to determine which variables are independent in a Bayes net”. Jessica Noss. EECS MIT. http://web.mit.edu/jmn/www/6.034/d-separation.pdf
- Chapter 10: “Directed graphical models”, pages 307-311 and pages 324-327 of Kevin Murphy’s book “Machine Learning: A Probabilistic Perspective”
- Koller, D., and Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.