CPSC 440: Advanced Machine Learning

Multivariate Gaussian

Mark Schmidt

University of British Columbia

Winter 2022

Bonus Slide Switch to Beamer

- Starting in this lecture, most slides will be in LaTeX.
- Why the change?
  - I have made major changes to the course this year (hopefully improvements).
  - But it is hard to prepare three 50-minute lectures per week.
  - So am going to rely much more on my old material.
- I am going to try to put the old material into the "story" of the current course.
  - But material was aimed at grad students who do a lot of "filling in the blanks".
    - Slow me down if I am going way too fast.
- Notation in these slides will be the same, but bonus slides will be this colour.
  - And Beamer slides do not work quite as well for annotation (you will see why).

Product of Gaussians in Matrix Notation

- If we have d variables, we could make each follow an independent Gaussian,

$$x_j^i \sim \mathcal{N}(\mu_j, \sigma_j^2),$$

- In this case the joint density $p(x^i \mid \mu_1, \mu_2, \dots, \mu_d, \sigma_1^2, \sigma_2^2, \dots, \sigma_d^2)$ can be written:

$$\begin{aligned}
\prod_{j=1}^{d} p(x_j^i \mid \mu_j, \sigma_j^2) &\propto \prod_{j=1}^{d} \exp\left(-\frac{(x_j^i - \mu_j)^2}{2\sigma_j^2}\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{d}\frac{1}{\sigma_j^2}(x_j^i - \mu_j)^2\right) && (e^a e^b = e^{a+b}) \\
&= \exp\left(-\frac{1}{2}(x^i - \mu)^T \Sigma^{-1}(x^i - \mu)\right) && \text{(matrix notation)}
\end{aligned}$$

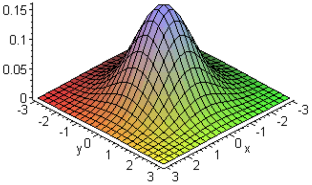
where $\mu = (\mu_1, \mu_2, \dots, \mu_d)$ and Σ is a diagonal matrix with diagonal elements σ_j^2.

- Distributions with this form are a special case of the multivariate Gaussian.

Multivariate Gaussian Distribution

- A $d > 1$ generalization of unvariate Gaussian is the multivariate normal/Gaussian,

Bivariate Normal



http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html

- This maintains many of the nice properties of univariate Gaussians.
  - Closed-form intuitive MLE, many analytic properties, makes theory easier.

- Multivariate Gaussians with non-diagonal covariance Σ models correlations.
  - Can take into account that "adjacent rooms have similar values".

Multivariate Gaussian Distribution

- The probability density for the multivariate Gaussian is given by

$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right), \quad \text{or } x^i \sim \mathcal{N}(\mu,\Sigma),$$

  where $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric with $\Sigma \succ 0$, and $|\Sigma|$ is the determinant.
  - Writing $\Sigma \succ 0$ means eigenvalues of Σ are all positive (Σ is "positive definite").
    - It does require that all elements of Σ are positive.
  - Or equivalently that $v^T\Sigma v > 0$ for all vectors $v \neq 0$ (implies Σ is invertible).

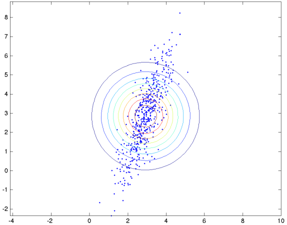
- Where does this wonky formula come from?
  - Consider a product of independent Gaussians, $z_j^i \sim \mathcal{N}(0,1)$.
  - Then perform a linear transformation, $x^i = Az^i + \mu$.
    - If we define $\Sigma = AA^T$, multivariate Gaussian is PDF of transformed variables.
    - Derivation in bonus slides.

- If $|\Sigma| = 0$ we say the Gaussian is degenerate (bonus).
  - Transformed variables x^i don't span the full space.

Multivariate Gaussian and Product of Gaussians

- The effect of a diagonal Σ on the multivariate Gaussian:
  - If $\Sigma = \alpha I$ the level curves are circles: 1 parameter.
  - If $\Sigma = D$ (diagonal) then axis-aligned ellipses: d parameters.
    - We saw that this is equivalent to using a product of independent Gaussians.
  - If Σ is dense they do not need to be axis-aligned: $d(d+1)/2$ parameters.

(by symmetry, we only need upper-triangular part of Σ)



- Diagonal Σ assumes features are independent, dense Σ models dependencies.

Independence in Gaussians

- Independence in multivariate Gaussian:
  - Independence between pairs of x_j is determined by covariance off-diagonals:

$$x_i \perp x_j \Leftrightarrow \Sigma_{ij} = 0,$$

    (so if Σ is diagonal the x_j are mutually independent).

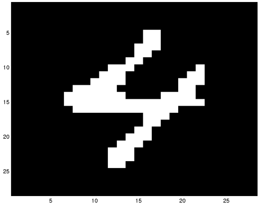
- If we allow Σ_{ij} to be non-zero, it models correlation between x_i and x_j.
  - We will see mathematically how the covariance relates to independence shortly.
  - This correlation can be positive or negative.

- Multivariate Gaussian is different than previous "product of whatever" models.
  - Multivariate Gaussian can model dependencies between all pairs of variables.
  - But, Gaussians do not directly model dependencies between triplets.
    - Or other higher-order interactions.

Example: Multivariate Gaussians on Digits

- Recall the task of density estimation with handwritten images of digits:

$$x^i = \text{vec}\left(\right. $$

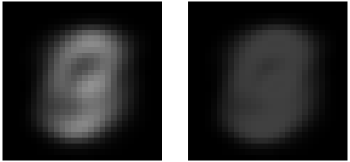


$$\left. \right),$$

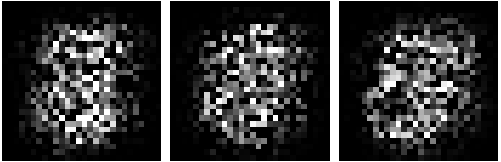
- Let's treat this as a continuous density estimation problem.

Example: Multivariate Gaussians on Digits

- MLE of parameters using independent Gaussians (diagonal Σ):
  - Mean μ_j (left) and variance σ_j^2 (right) for each feature.



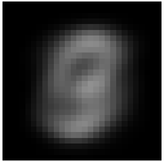
- Samples generate from this model:



- Because Σ is diagonal, doesn't model dependencies between pixels.

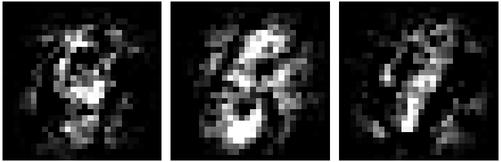
Example: Multivariate Gaussians on Digits

- MLE of mean vector using multivariate Gaussians (dense Σ):



  - Which is the same as diagonal case (784×784 covariance not shown).

- Samples generate from this model:



- Captures pairwise correlations between pixels, but only between pairs.

Outline

Inference in Multivariate Gaussian

- How do we do predictions/inference in the model?
  - We can compute likelihood of data $p(x)$ by plugging into formula.
    - As with univariate variate Gaussian, likelihood is not a probability.
  - The decoding of the vector x is given by the mean μ.
  - But what about deriving marginals like $p(x_j)$?
    - You could use marginals to compute probability that x_j falls in an interval.
  - Or computing conditionals like $p(x_j \mid x_{j'})$?
    - Maybe you know the values of some variables and want to "fill in" others.
  - Or generating samples from the distribution (for Monte Carlo inference)?

- Gaussians have many nice properties that make many computations easy.
  - Rather than giving a list of properties, we will introduce them "as needed".
  - A multivariate Gaussian "cheat sheet" is here:
    - https://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gaussians.pdf
  - For a more-careful discussion of Gaussians, see the playlist here:
    - https://www.youtube.com/watch?v=TC0ZAX3DA88&t=2s&list=PL17567A1A3F5DB5E4&index=34

Affine Property of Gaussians: Special Case of Shift

- Assume that random variable x follows a Gaussian distribution,

$$z \sim \mathcal{N}(\mu, \Sigma).$$

- And consider shifting the random variable by a vector b,

$$x = z + b.$$

- Then random variable x follows a Gaussian distribution

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

where we've shifted the mean.

Affine Property of Gaussians: General Case

- Assume that random variable x follows a Gaussian distribution,

$$z \sim \mathcal{N}(\mu, \Sigma).$$

- And consider an affine transformation of the random variable,

$$x = Az + b.$$

- Then random variable x follows a Gaussian distribution

$$x \sim \mathcal{N}(A\mu + b, A\Sigma A^\top),$$

although note we might have $|A\Sigma A^\top| = 0$.
  - For example, if x has a higher-dimension that z.

Sampling from a Multivariate Gaussian

- The affine property of multivariate Gaussian:
  - If $z \sim \mathcal{N}(\mu, \Sigma)$, then $Az + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.

- To sample from a general multivariate Gaussian $\mathcal{N}(m, C)$:
  1. Sample z from a $\mathcal{N}(0, I)$.
     - Each z_j comes independently from the "standard normal" $\mathcal{N}(0, 1)$.
  2. Transform z to a sample from the right Gaussian using the affine property:

$$Az + m \sim \mathcal{N}(m, \underbrace{AA^T}_{C}),$$

  where we choose A so that $AA^T = C$.
  - One way to compute A from C is the Cholesky factorization (`cholesky` in Julia).

Inference Task: Marginalization

- Consider the inference task of marginalization.
  - Going from the joint $p(x_1, x_2, \ldots, x_d)$ to the marginal $p(x_j)$.

- We can do this with the marginalization rule,

$$p(x_j) = \int_{x_1} \cdots \int_{x_{j-1}} \int_{x_{j+1}} \cdots \int_{x_d} p(x \mid \mu, \Sigma) dx_d \cdots x_{j+1} dx_{j-1} \cdots dx_1,$$

  but this integral may be unpleasant.

- For Gaussians, the affine property allows us to easily derive the marginal.

Partitioned Gaussian

- Consider a dataset where we've partitioned the variables into two sets:

$$X = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & z_1 & z_2 \\ | & | & | & | \end{bmatrix}.$$

- It's common to write multivariate Gaussian for partitioned data as:

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right),$$

- Example:

$$\text{If } \begin{bmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0.3 \\ -0.1 \\ 1.5 \\ 2.5 \end{bmatrix}, \begin{bmatrix} 1.5 & -0.1 & -0.1 & 0 \\ -0.1 & 2.3 & 0.1 & 0 \\ -0.1 & 0.1 & 1.6 & -0.2 \\ 0 & 0 & -0.2 & 4 \end{bmatrix}\right), \quad \text{then} \quad \mu_z = \begin{bmatrix} 1.5 \\ 2.5 \end{bmatrix} \quad \text{and} \quad \Sigma_{zz} = \begin{bmatrix} 1.6 & -0.2 \\ -0.2 & 4 \end{bmatrix}.$$

- The blocks do not have to be the same size.

Marginalization of Gaussians

- Consider a dataset where we've partitioned the variables into two sets:

$$X = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & z_1 & z_2 \\ | & | & | & | \end{bmatrix}.$$

- It's common to write multivariate Gaussian for partitioned data as:

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right),$$

- If I want the marginal distribution $p(x)$, I can use the affine property,

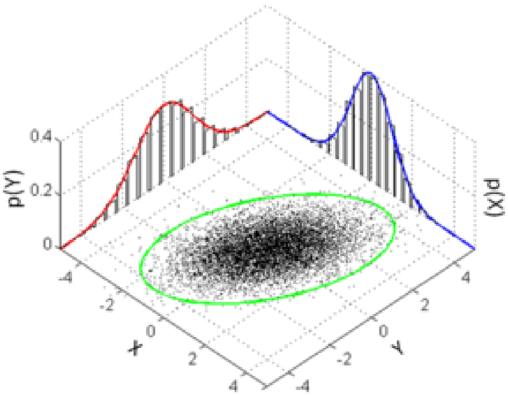
$$x = \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{A} \begin{bmatrix} x \\ z \end{bmatrix} + \underbrace{0}_{b},$$

  to get that

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$

Marginalization of Gaussians

- In a picture, ignoring a subset of the variables gives a Gaussian:



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Conditioning in Gaussians

- Again consider a partitioned Gaussian,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right).$$

- Using a lot linear algebra (see textbook), conditional probabilities are Gaussian,

$$x \mid z \sim \mathcal{N}(\mu_{x \mid z}, \Sigma_{x \mid z}),$$

  where

$$\mu_{x \mid z} = \mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(z - \mu_z), \quad \Sigma_{x \mid z} = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}.$$

- "For any fixed z, the distribution of x is a Gaussian".
  - Notice that if $\Sigma_{xz} = 0$ then x and z are independent ($\mu_{x \mid z} = \mu_x$, $\Sigma_{x \mid z} = \Sigma_{xx}$).
    - Since if $\Sigma_{xz} = 0$ we have $p(x \mid z) = p(x)$.

Conditional Independence in Gaussians

- Independence in Gaussians is determined by sparsity pattern of the covariance Σ.
  - Sparsity pattern: "where the non-zeroes are".

- Conditional independence in Gaussians is determined by inverse of covariance Σ.
  - We call the inverse the precision matrix Θ, so $\Theta \triangleq \Sigma^{-1}$.
  - Specifically, conditional independence is determined by the sparsity pattern of Θ.

- We use the sparsity pattern of Θ to define a graph.
  - Each node in the graph corresponds to a variable $j \in \{1, 2, \ldots, d\}$.
  - Each edge in the graph corresponds to a non-zero Θ_{ij}.

- Checking independence and conditional independence using the graph:
  - $x_i \perp x_j$ if no path exists between x_i and x_j in the graph.
  - $x_i \perp x_j \mid x_k$ if x_k blocks all paths from x_i to x_j in the graph.
    - Technically, this only checks whether independence is implied by the sparsity pattern.

Conditional Independence in Gaussians

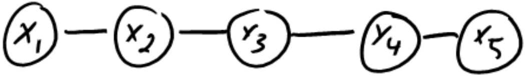
- Consider a Gaussian with the following covariance matrix:

$$\Sigma = \begin{bmatrix} 0.0494 & -0.0444 & -0.0312 & 0.0034 & -0.0010 \\ -0.0444 & 0.1083 & 0.0761 & -0.0083 & 0.0025 \\ -0.0312 & 0.0761 & 0.1872 & -0.0204 & 0.0062 \\ 0.0034 & -0.0083 & -0.0204 & 0.0528 & -0.0159 \\ -0.0010 & 0.0025 & 0.0062 & -0.0159 & 0.2636 \end{bmatrix}$$

- $\Sigma_{ij} \neq 0$ so all variables are dependent: $x_1 \not\perp x_2$, $x_1 \not\perp x_5$, and so on.
  - This would show up in graph: you would be able to reach any x_i from any x_j.
- The inverse is given by a tri-diagonal matrix:

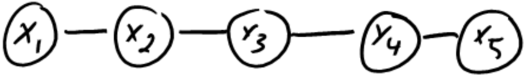
$$\Sigma^{-1} = \begin{bmatrix} 32.0897 & 13.1740 & 0 & 0 & 0 \\ 13.1740 & 18.3444 & -5.2602 & 0 & 0 \\ 0 & -5.2602 & 7.7173 & 2.1597 & 0 \\ 0 & 0 & 2.1597 & 20.1232 & 1.1670 \\ 0 & 0 & 0 & 1.1670 & 3.8644 \end{bmatrix}$$

- So conditional independence is described by a 5-node "chain'-structured" graph:



Conditional Independence in Gaussians

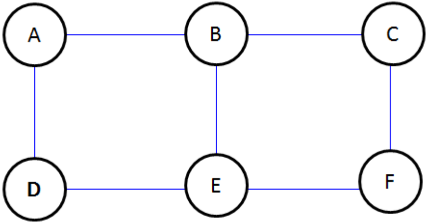
- All variables are dependent in this graph, since a path exists.

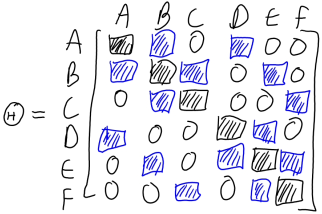


- But we have many conditional independences such as:
  - $x_1 \perp x_3 \mid x_2$.
  - $x_2 \perp x_5 \mid x_4$.
  - $x_1 \perp x_5 \mid x_3$.
  - $x_1 \perp x_3, x_4, x_5 \mid x_2$ (we will later call this specific one the "Markov property").
  - $x_1, x_2 \perp x_4, x_5 \mid x_3$.

Conditional Independence in Gaussian

- Checking conditional independence among variable groups in Gaussians:
  - $A \perp B \mid C$ if C blocks all paths from any A to any B.





- Example:
  - $A \not\perp C$.
  - $A \not\perp C \mid B$.
  - $A \perp C \mid B, E$.
  - $A, B \not\perp F \mid C$
  - $A, B \perp F \mid C, E$.

Summary

- Multivariate Gaussian generalizes univariate Gaussian for multiple variables.
  - Parameterized by mean vector μ and positive-definite covariance matrix Σ.
  - Product of independent Gaussians is equivalent to using a diagonal Σ.
  - Models correlations between paris of variables with non-zero off-diagonals in Σ.
- Inference multivariate Gaussian:
  - Affine transformations of Gaussians are Gaussians (can be used to sample).
  - Marginals and conditionals of Gaussians are Gaussians.
- Conditional independence in multivariate Gaussians:
  - Precision matrix Θ is inverse of Σ.
  - Conditional independence determined by off-diagonals in Θ.
  - We use the non-zero off-diagonals in Θ to define a graph.
  - Variables are independent if all paths are blocked by conditioning variables.

- Next time: learning the graph?

Positive-Definiteness of Θ and Checking Positive-Definiteness

- If we define centered vectors $\tilde{x}^i = x^i - \mu$ then empirical covariance is

$$S = \frac{1}{n}\sum_{i=1}^n (x^i - \mu)(x^i - \mu)^\top = \frac{1}{n}\sum_{i=1}^n \tilde{x}^i(\tilde{x}^i)^\top = \frac{1}{n}\tilde{X}^\top\tilde{X} \succeq 0,$$

  so S is positive semi-definite but not positive-definite by construction.
- If data has noise, it will be positive-definite with n large enough.
- For $\Theta \succ 0$, note that for an upper-triangular T we have

$$\log|T| = \log(\text{prod}(\text{eig}(T))) = \log(\text{prod}(\text{diag}(T))) = \text{Tr}(\log(\text{diag}(T))),$$

  where we've used Matlab notation.
- So to compute $\log|\Theta|$ for $\Theta \succ 0$, use Cholesky to turn into upper-triangular.
  - Bonus: Cholesky fails if $\Theta \succ 0$ is not true, so it checks positive-definite constraint.

Positive-Definite implies Invertibility

- If $A \succ 0$, then all the eigenvalues of A are positive.
- If each eigenvalue is positive, the product of the eigenvalues is positive.
- The product of the eigenvalues is equal to the determinant.
- Thus, the determinant is positive.
- The determinant not being 0 implies the matrix is invertible.

Multivariate Gaussian from Univariate Gaussians

- Consider a joint distribution that is the product univariate standard normals:

$$p(z^i) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_j^i)^2\right)$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\langle z^i, z^i\rangle\right).$$

- Now define $x^i = Az^i + \mu$ for some (non-singular) matrix A and vector μ.
- The change of variables formula for multivariate probabilities is

$$p(x^i) = p(z^i)\left|\frac{\partial z^i}{\partial x^i}\right|.$$

- Plug in $z^i = A^{-1}(x^i - \mu)$ and $\frac{\partial z^i}{\partial x^i} = A^{-1}$...

Multivariate Gaussian from Univariate Gaussians

- This gives

$$p(x^i \mid \mu, A) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(\frac{1}{2}\langle A^{-1}(x^i - \mu), A^{-1}(x^i \mu)\rangle\right) |\det(A^{-1})|$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}} |\det(A)|} \exp\left(\frac{1}{2}(x^i - \mu)A^{-\top}A^{-1}(x^i - \mu)\right).$$

- Define $\Sigma = AA^\top$ (so $\Sigma^{-1} = A^{-\top}A^{-1}$ and $\det \Sigma = (\det A)^2$) to get

$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu)\right)$$

- So multivariate Gaussian is an affine transformtation of independent Gaussians.

Degenerate Gaussians

- If $|\Sigma| = 0$, we say the Gaussian is degenerate.
- In this case the PDF only integrates to 1 along a subspace of the original space.
- With $d = 2$ degenerate Gaussians only have non-zero probability along a line (or just one point).