# CPSC 440: Advanced Machine Learning

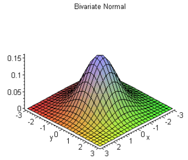
## More Gaussians

Mark Schmidt

University of British Columbia

Winter 2021

# Last Time: Multivariate Gaussian




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

- The multivariate normal/Gaussian distribution models PDF of vector $x^i$ as

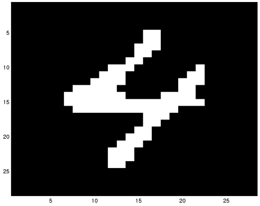
$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu)\right)$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma \succ 0$.
  - This is the density for a linear transformation of a product of independent Gaussians.
- MLE is easy: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x^i$, and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x^i - \mu)(x^i - \mu)^\top$.
- Diagonal $\Sigma$ implies independence between variables.

# Example: Multivariate Gaussians on Digits

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

$$x^i = \text{vec}\Bigg($$

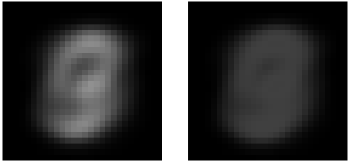



$$\Bigg),$$

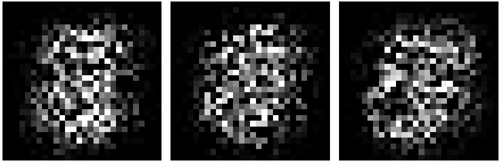
- Let's treat this as a continuous density estimation problem.

# Example: Multivariate Gaussians on Digits

- MLE of parameters using independent Gaussians (diagonal $\Sigma$):
  - Mean $\mu_j$ (left) and variance $\sigma_j^2$ (right) for each feature.



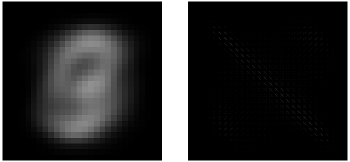
- Samples generate from this model:



- Because $\Sigma$ is diagonal, doesn't model dependencies between pixels.

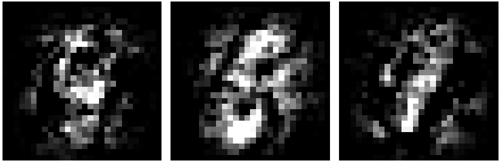
# Example: Multivariate Gaussians on Digits

- MLE of parameters using multivariate Gaussians (dense $d \times d$ covariance $\Sigma$):



  - Largest values are on main diagonal (self-correlation), above/below main diagonal (neighbour above/below in image), and shifted (neighbour left/right in image).

- Samples generate from this model:



- Captures some pairwise dependencies between pixels, but not expressive enough.

# MAP Estimation in Multivariate Gaussian (Trace Regularization)

- A classic regularizer for $\Sigma$ is to add a diagonal matrix to $S$ and use

$$\Sigma = S + \lambda I,$$

which satisfies $\Sigma \succ 0$ because $S \succeq 0$ (eigenvalues at least $\lambda$).

- This corresponds to L1-regularization of diagonals of precision.

$$\begin{aligned}
f(\Theta) &= \text{Tr}(S\Theta) - \log|\Theta| + \lambda \sum_{j=1}^{d} |\Theta_{jj}| && \text{(Gauss. NLL plus L1 of diags)}\\
&= \text{Tr}(S\Theta) - \log|\Theta| + \lambda \sum_{j=1}^{d} \Theta_{jj} && \text{(Diagonals of pos. def. matrix are } > 0)\\
&= \text{Tr}(S\Theta) - \log|\Theta| + \lambda \text{Tr}(\Theta) && \text{(Definition of trace)}\\
&= \text{Tr}(S\Theta + \lambda\Theta) - \log|\Theta| && \text{(Linearity of trace)}\\
&= \text{Tr}((S + \lambda I)\Theta) - \log|\Theta| && \text{(Distributive law)}
\end{aligned}$$

- Taking gradient and setting to zero gives $\Sigma = S + \lambda$.
  - But doesn't set to exactly zero as log-determinant term is too "steep" at 0.

# Graphical LASSO

- A popular generalization called the graphical LASSO,

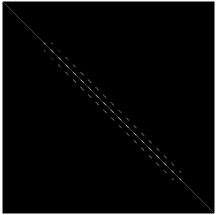
$$f(\Theta) = \text{Tr}(S\Theta) - \log|\Theta| + \lambda\|\Theta\|_1.$$

where we are using the element-wise L1-norm, $\|\Theta\|_1 = \sum_{i=1}^d \sum_{j=1}^d \Theta_{ij}$.

- Gives sparse off-diagonals in $\Theta$.
  - Can solve very large instances with proximal-Newton and other tricks ("QUIC").
- It's common to draw the non-zeroes in $\Theta$ as a graph.
  - Has an interpretation in terms on conditional independence (we'll cover this later).

# Graphical LASSO on Digits

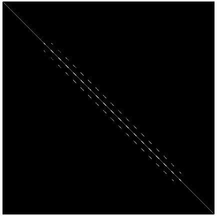
- Sparsity pattern if we instead use the graphical LASSO:
  - MAP estimate of precision matrix $\Theta$ with regularizer $\lambda\|\Theta\|_1$ (with $\lambda = 1/8$).



- To understand this picture, first consider the two matrices:
  - The images of digits, which are $m \times m$ matrices ($m$ pixels by $m$ pixels)
    - This gives $d = m^2$ elements of $x^i$, which we'll assume are in "column-major" order.
    - So the first $m$ elements of $x^i$ are row 1, the next $m$ elements are row 2, and so on.
  - The covariance picture above, which is $d \times d$ so will be $m^2 \times m^2$.

# Graphical LASSO on Digits

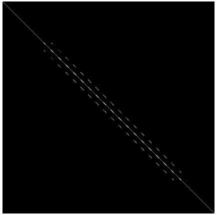
- Sparsity pattern if we instead use the graphical LASSO:
  - MAP estimate of precision matrix $\Theta$ with regularizer $\lambda\|\Theta\|_1$ (with $\lambda = 1/8$).



- So what are the non-zeroes in the covariance matrix?
  1. The diagonals $\Theta_{i,i}$ (these are all non-zero because $\Theta \succ 0$).
  2. The first off-diagonals $\Theta_{i,i+1}$ and $\Theta_{i+1,i}$.
     - This represents the dependencies between adjacent pixels horizontally.
  3. The $(m+1)$ off-diagonals $\Theta_{i,i+m}$ and $\Theta_{i+m,i}$.
     - This represents the dependencies between adjacent pixels vertically.
     - Because in "column-major" order, you go "down" a pixel every $m$ indices.

# Graphical LASSO on Digits

- Sparsity pattern if we instead use the graphical LASSO:
  - MAP estimate of precision matrix $\Theta$ with regularizer $\lambda\|\Theta\|_1$ (with $\lambda = 1/8$).



- The graph represented by this adjacency matrix is (roughly) the 2d image lattice.
  - Pixels that are near each other in the image end up being connected by an edge.

- Examples:
  - https://normaldeviate.wordpress.com/2012/09/17/high-dimensional-undirected-graphical-models

# Outline

# Inference in Multivariate Gaussian

- Suppose we have fit $\mu$ and $\Sigma$ to our data $X$.
  - Using either MLE or MAP.

- How do we do predictions/inference in the model?
  - We can compute likelihood of data $p(x)$ by plugging into formula.
    - Likelihood of seeing the vector $x$?
  - But what about computing a marginal likelihood like $p(x_j)$?
    - What is the likelihood that variable $j$ takes the value $x_j$?
  - Or computing a conditional likelihood $p(x_j \mid x_{j'})$.
    - Maybe you know the values of some variables and want to "fill in" others.
  - Or generating samples from the distribution.

- Gaussians have many nice properties that make these computations easy.

# Closedness of Multivariate Gaussian

- Multivariate Gaussian has nice properties of univariate Gaussian:
  - Closed-form MLE for $\mu$ and $\Sigma$ given by sample mean/variance.
  - Central limit theorem: mean estimates of random variables converge to Gaussians.
  - Maximizes entropy subject to fitting mean and covariance of data.

- A crucial computational property: Gaussians are closed under many operations.
  1. Affine transformation: if $p(x)$ is Gaussian, then $p(Ax + b)$ is a Gaussian[1].
  2. Marginalization: if $p(x, z)$ is Gaussian, then $p(x)$ is Gaussian.
  3. Conditioning: if $p(x, z)$ is Gaussian, then $p(x \mid z)$ is Gaussian.
  4. Product: if $p(x)$ and $p(z)$ are Gaussian, then $p(x)p(z)$ is proportional to a Gaussian.

- Most continuous distributions don't have these nice properties.

[1] Could be degenerate with $|\Sigma| = 0$, dependending on particular $A$.

# Affine Property: Special Case of Shift

- Assume that random variable $x$ follows a Gaussian distribution,

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

- And consider an shift of the random variable,

$$z = x + b.$$

- Then random variable $z$ follows a Gaussian distribution

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

where we've shifted the mean.

# Affine Property: General Case

- Assume that random variable $x$ follows a Gaussian distribution,

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

- And consider an affine transformation of the random variable,

$$z = Ax + b.$$

- Then random variable $z$ follows a Gaussian distribution

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

  although note we might have $|A\Sigma A^\top| = 0$.

# 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:

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

- The blocks don't necessarily have to have 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,

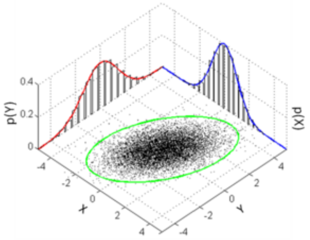
$$x = \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{A} \begin{bmatrix} x \\ z \end{bmatrix} + \underbrace{0}_{b},$$

  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

- This seems less intuitive if you use rules of probability to marginalize:

$$p(x) = \int_{z_1}\int_{z_2}\cdots\int_{z_d} \frac{1}{(2\pi)^{\frac{d}{2}}\left|\begin{bmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{bmatrix}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}x\\ z\end{bmatrix} - \begin{bmatrix}\mu_x\\ \mu_z\end{bmatrix}\right)\begin{bmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{bmatrix}^{-1}\left(\begin{bmatrix}x\\ z\end{bmatrix} - \begin{bmatrix}\mu_x\\ \mu_z\end{bmatrix}\right)\right) dz_d dz_{d-1}\ldots dz_1.$$

- A caution about different "precisions": note that $\Sigma_{xx}^{-1} \neq (\Sigma^{-1})_{xx}$ in general.

# Conditioning in Gaussians

- Again consider a partitioned Gaussian,

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

- The conditional probabilities are also Gaussian,

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

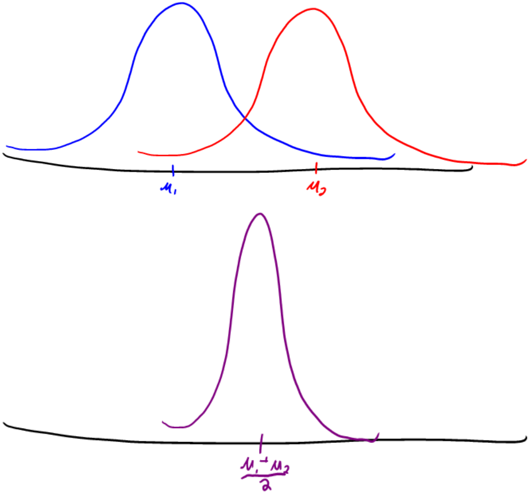
  where

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

- "For any fixed $z$, the distribution of $x$ is a Gaussian".
  - Notice that if $\Sigma_{xz} = 0$ then $x$ and $z$ are independent ($\mu_{x \mid z} = \mu_x$, $\Sigma_{x \mid z} = \Sigma_x$).
  - We previously saw the special case where $\Sigma$ is diagonal (all variables independent).

# Product of Gaussian Densities

- If $\Sigma_1 = I$ and $\Sigma_2 = I$ then product of PDFs has $\Sigma = \frac{1}{2}I$ and $\mu = \frac{\mu_1+\mu_2}{2}$.



# Product of Gaussian Densities

- Let $f_1(x)$ and $f_2(x)$ be Gaussian PDFs defined on variables $x$.
- The product of the PDFs $f_1(x)f_2(x)$ is proportional to a Gaussian density,
  - With $(\mu_1, \Sigma_1)$ as parameters of $f_1$ and $(\mu_2, \Sigma_2)$ for $f_2$:

$$\text{covariance of} \quad \Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}.$$

$$\text{mean of } \mu = \Sigma\Sigma_1^{-1}\mu_1 + \Sigma\Sigma_2^{-1}\mu_2,$$

  although this density may not be normalized (may not integrate to 1 over all $x$).
- So if we can write a probability as $p(x) \propto f_1(x)f_2(x)$ for 2 Gaussians, then $p$ is a Gaussian with known mean/covariance.

# Product of Gaussian Densities

- Example of a Gaussian likelihood $p(x^i \mid \mu, \Sigma)$ for IID data,

$$\prod_{i=1}^{n} p(x^i \mid, \mu, \Sigma),$$

will be Gaussian if the individual likelihoods $p(x^i \mid \mu, \Sigma)$ are Gaussian.

- Example of a Gaussian likelihood $p(x^i \mid \mu, \Sigma)$ and Gaussian prior $p(\mu \mid \mu_0, \Sigma_0)$.
  - Posterior for $\mu$ will be Gaussian:

$$\begin{aligned} p(\mu \mid x^i, \Sigma, \mu_0, \Sigma_0) &\propto p(x^i \mid \mu, \Sigma) p(\mu \mid \mu_0, \Sigma_0) && \text{(Bayes rule)} \\ &= p(\mu \mid x^i, \Sigma) p(\mu \mid \mu_0, \Sigma_0) && (\text{symmetry of } x^i \text{ and } \mu) \\ &= \text{(some Gaussian)}. \end{aligned}$$

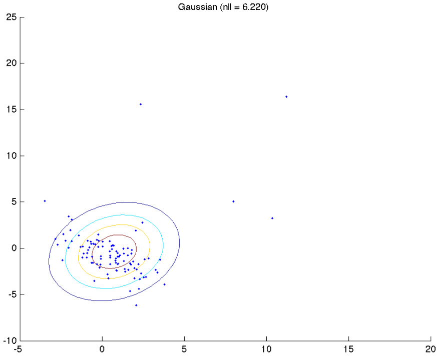
- Non-example of $p(x_2 \mid x_1)$ being Gaussian and $p(x_1 \mid x_2)$ being Gaussian.
  - Product $p(x_2 \mid x_1) p(x_1 \mid x_2)$ may not be a proper distribution.
  - Although we saw it will be a Gaussian if they are independent.
- "Product of Gaussian densities" will be used later in Gaussian Markov chains.

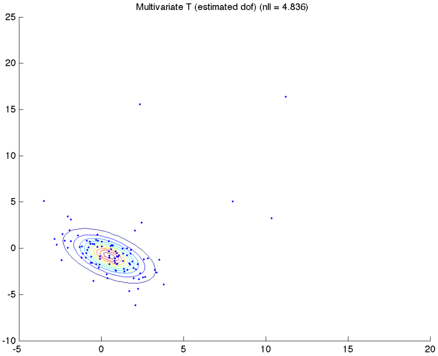
# Properties of Multivariate Gaussians

- A multivariate Gaussian "cheat sheet" is here:
  - https://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gaussians.pdf
- For a careful discussion of Gaussians, see the playlist here:
  - https://www.youtube.com/watch?v=TC0ZAX3DA88&t=2s&list=PL17567A1A3F5DB5E4&index=34

# Problems with Multivariate Gaussian

- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace or multivariate T.

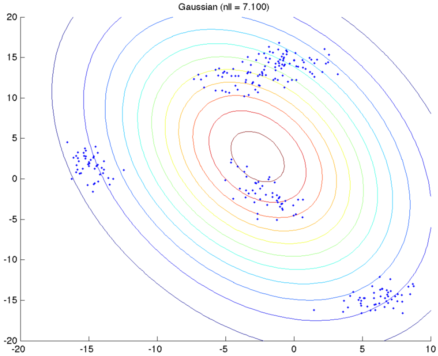



- These require numerical optimization to compute MLE/MAP.

# Problems with Multivariate Gaussian

- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace of multivariate T.
  - Still unimodal, which often leads to very poor fit.



# Summary

- MAP in multivariate Gaussian:
  - Common approach is trace regularization, graphical Lasso gives visualization.
- Properties of multivariate Gaussian:
  - Closed under affine transformations, marginalization, conditioning, and products.
  - But unimodal and not robust.
- Next time: a universal model for continuous densities.

# MAP for Univariate Gaussian Mean

- Assume $x^i \sim \mathcal{N}(\mu, \sigma^2)$ and assume $\mu \sim \mathcal{N}(\mu_0, 1)$.
- The MAP estimate of $\mu$ under these assumptions can be written as

$$\hat{\mu} = \frac{n}{n+\sigma^2}\bar{x} + \frac{\sigma^2}{n+\sigma^2}\mu_0,$$

  where $\bar{x}$ is the sample mean, $\frac{1}{n}\sum_{i=1}^{n} x^i$ (which is the MLE).
- The MAP estimate is a convex combination of the MLE and prior mean $\mu_0$.
  - Regularizer moves us in a straight line away from MLE towards $\mu_0$.
  - With small $n$ you stay close to prior, with large $n$ you start ignoring prior.