Multivariate Gaussians

CPSC 440/550: Advanced Machine Learning

cs.ubc.ca/~dsuth/440/23w2

University of British Columbia, on unceded Musqueam land

2023-24 Winter Term 2 (Jan–Apr 2024)

Last time: Univariate Gaussians, Bayesian learning

- Continuous density estimation with the Gaussian=normal distribution

$$x \sim \mathcal{N}(\mu, \sigma^2) \quad \text{means} \quad p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Cumulative distribution function (cdf) $F(t)$
- Inverse probability sampling: $F^{-1}(U)$ for $U \sim \mathrm{Unif}([0,1])$
- MLE: sample mean, sample variance (with the $1/n$)
- With fixed variance: conjugate prior for the mean is Gaussian

- Gaussian likelihood gives linear regression/square loss; MAP gives ridge regression
- Bayesian learning integrates over model uncertainty
  - Posterior predictive: $p(\tilde{y} \mid \tilde{x}, \mathbf{X}, \mathbf{y}) = \int p(\tilde{y} \mid w) p(w \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}w$
  - Beta-Bernoulli model: use posterior $\mathrm{Beta}(n_1 + \alpha, n_0 + \beta)$

Bayesian learning in the Categorical-Dirichlet model

- If $X \mid \boldsymbol{\theta} \sim \mathrm{Cat}(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \mid \boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, we saw before that

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha}) \propto p(\mathbf{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) &\propto \theta_1^{n_1} \cdots \theta_k^{n_k} \theta_1^{\alpha_1 - 1} \cdots \theta_1^{n_k - 1} \\
&= \theta_1^{(n_1+\alpha_1)-1} \cdots \theta_k^{(n_k+\alpha_k)-1}
\end{aligned}
$$

$$
\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha} \sim \mathrm{Dir}(\mathbf{n} + \boldsymbol{\alpha}) \qquad \text{where } \mathbf{n} \in \mathbb{R}^d,\ n_j = \sum_{i=1}^{n} \mathbb{1}\left(x^{(i)} = j\right)
$$

- MAP: $$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\theta \mid \mathbf{X}) \propto \mathbf{n} + \boldsymbol{\alpha} - 1$$

- Bayesian learning uses the posterior predictive distribution,

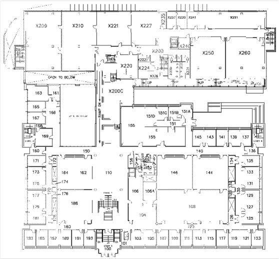
$$
\begin{aligned}
p(x = c \mid \mathbf{X}, \boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} p(x = c \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \theta_c\, p(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\theta} \quad = \mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \mathrm{Dir}(\mathbf{n}+\boldsymbol{\alpha})} [\theta_c] \quad \propto \mathbf{n} + \boldsymbol{\alpha}
\end{aligned}
$$

Multivariate Gaussian

- To handle Bayesian linear regression, we're going to need one more tool: multivariate Gaussians
  - (Also useful much more broadly . . .)

Motivating problem: Measuring building air quality

- Want to measure "air quality" across rooms in a building
- Measure pollutant concentrations (PM10, CO, O3, ...) in each room over time:



Rm 1	Rm 2	Rm 3	Rm 4	Rm 5	Rm 6	Rm 7	Rm 8	Rm 9
0.1	1.4	0.2	1.8	1.0	1.0	0.1	0.1	1.1
0.2	1.3	0.1	1.9	1.1	0.9	0.1	0.1	1.1
0.1	0.3	1.4	2.0	0.7	0.3	0.1	0.2	0.4
0.1	1.1	0.2	2.1	1.1	1.1	0.1	0.3	0.5
2.7	2.6	2.5	5.1	2.4	2.8	3.2	2.5	3.1
0.1	0.4	0.2	1.8	1.3	0.4	0.1	0.4	1.0
0.1	1.2	0.2	1.8	1.4	1.1	0.7	0.7	0.5

- We can model this data to identify patterns/problems:
  - Some rooms usually have worse air than others
  - Some rooms' quality may be correlated with others' (adjacent, shared air...)
  - Also temporal correlations, which we won't handle yet

Start: product of Gaussians

- Like before, simplest thing to do is to make different dimensions independent

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

- Gives joint density

$$p(x \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{d} p(x_j \mid \mu_j, \sigma_j^2) \propto \prod_{j=1}^{d} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

$$= \exp\left(-\frac{1}{2}\sum_{j=1}^{d} \frac{(x_j - \mu_j)^2}{\sigma_j^2}\right) = \exp\left(-\frac{1}{2}(x - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})\right)$$

where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_j^2 \end{bmatrix}$

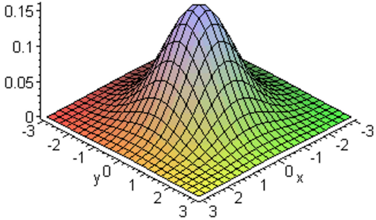
Multivariate Gaussians

- General multivariate Gaussian: $\boldsymbol{\Sigma}$ doesn't have to be diagonal

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

  - $|\boldsymbol{\Sigma}|$ is the determinant (product of eigenvalues)

- Many nice properties, like univariate case
  - Closed-form, intuitive MLE
  - Conjugate priors
  - Many nice analytic properties
  - Multivariate central limit theorem
  - ...



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

- Off-diagonal covariance entries give covariance: $\operatorname{Cov}(x_j, x_{j'}) = \Sigma_{jj'}$
  - "Adjacent rooms have similar air qualities"
  - Correlation is $\operatorname{Cov}(x_j, x_{j'}) / \sqrt{\operatorname{Var}(x_j) \operatorname{Var}(x_{j'})} = \Sigma_{jj'} / \sqrt{\Sigma_{jj} \Sigma_{j'j'}}$

Covariance matrices

- The $d \times d$ matrix $\mathbf{\Sigma}$ is called the covariance matrix, $\mathrm{Cov}(x)$
  - Also called "variance-covariance matrix"; sometimes written $\mathrm{Var}(x)$
- For *any* continuous distribution, $\mathrm{Var}(x) > 0$. What about multivariate dists?
- Consider the univariate random variable $v^\mathsf{T} x$. We have

$$\mathrm{Var}(v^\mathsf{T} x) = \mathrm{Var}\left(\sum_{j=1}^{d} v_j x_j\right) = \sum_{j=1}^{d}\sum_{j'=1}^{d} \mathrm{Cov}\left(v_j x_j, v_{j'} x_{j'}\right)$$

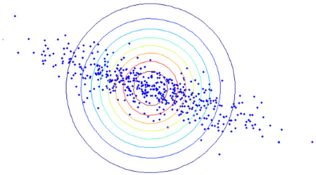
$$= \sum_{j=1}^{d}\sum_{j'=1}^{d} v_j \,\mathrm{Cov}\left(x_j, x_{j'}\right) v_{j'} = v^\mathsf{T} \mathbf{\Sigma} v$$

- A continuous multivariate random variable requires $v^\mathsf{T} \mathbf{\Sigma} v > 0$ for *all* v
- This is exactly the condition that $\mathbf{\Sigma}$ is strictly positive-definite
- Equivalent condition (see notes on website): all eigenvalues are positive
- Equivalent condition: there is some (full-rank) $A \in \mathbb{R}^{n \times n}$ such that $\mathbf{\Sigma} = A A^\mathsf{T}$

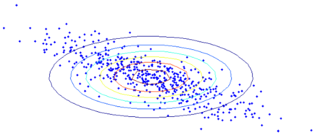
Kinds of covariances

- If $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$, level sets of the density are circles
  - One parameter
  - The $x_j \sim \mathcal{N}(0, \sigma^2)$ are mutually independent, because

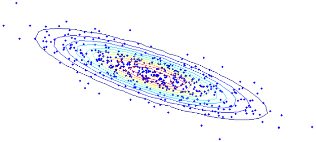
$$p(x \mid \sigma^2) = p(x_1 \mid \sigma^2) \cdots p(x_d \mid \sigma^2)$$



- If $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is diagonal: axis-aligned ellipses
  - d parameters
  - Each $x_j \sim \mathcal{N}(0, \sigma_j^2)$ is still independent



- For general $\mathbf{\Sigma}$, might not be axis-aligned
  - $d(d+1)/2$ parameters – not d^2 since $\mathbf{\Sigma}$ is symmetric
  - x_j can now be correlated



Degenerate Gaussians

- If $\mathbf{\Sigma} \succeq 0$ but not $\succ 0$ – it has some zero eigenvalues – we call it degenerate
- Means that there's some direction v where $v^\mathsf{T} \mathbf{\Sigma} v = 0$, i.e. $v^\mathsf{T} x$ is constant
- Standard density function doesn't exist (no inverse, i.e. divide-by-zero error)

- For $d = 1$, $\mathcal{N}(\mu, 0)$ is a point mass: every sample is exactly μ
- For $d = 2$, can be a point mass, or all samples can live along a line

- In general, has support on a subspace of dimension $\operatorname{rank} \mathbf{\Sigma}$
  - Has a Gaussian density with respect to that subspace

Affine transformations

- For any random vector x, we have that

$$\mathbb{E}[Ax + \mu] = A\,\mathbb{E}[x] + \mu$$
$$\operatorname{Cov}(Ax + \mu) = A\operatorname{Cov}(x)A^\mathsf{T}$$

- Fact (won't prove here; straightforward if you use characteristic functions): affine transformations of multivariate normals are multivariate normal
- So, if $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $Ax + b \sim \mathcal{N}(A\boldsymbol{\mu} + b, A\boldsymbol{\Sigma}A^\mathsf{T})$
- Even if x is non-degenerate, $A\boldsymbol{\Sigma}A^\mathsf{T}$ might be singular!
  - Examples: $A = 0$, or if x is one-dimensional and A is 5×1 ...

- This immediately gives us a nice sampling algorithm:
  - Sample d independent standard normals, $z_j \sim \mathcal{N}(0, 1)$
  - Return $AZ + \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, AA^\mathsf{T})$
    - Need to find an A such that $AA^\mathsf{T} = \boldsymbol{\Sigma}$
    - Can use Cholesky factorization (`np.linalg.cholesky`) to find a (lower-triangular) A
    - Or (a little slower), eigendecompose $\boldsymbol{\Sigma}$ and use $A^{\frac{1}{2}} = \sum_j \sqrt{\lambda_j} v_j v_j^\mathsf{T}$

Marginalizing Gaussians

- If we have a joint distribution over $x = (x_1, \ldots, x_d)$, might care about just x_j
- $p(x_j) = \int \cdots \int p(x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,\mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \cdots \mathrm{d}x_d$
- ...but we can skip that nasty integral by just thinking a little bit!
- Let's partition our variables into block matrices, $\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^\mathsf{T} & \boldsymbol{\Sigma}_z \end{bmatrix} \right)$
- For example,

$$\begin{bmatrix} x_1 \\ x_2 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0.6 \\ -1.3 \\ 9.8 \\ 0.1 \\ -3 \end{bmatrix}, \begin{bmatrix} 1.3 & -0.1 & -0.2 & 0.4 & 0 \\ -0.1 & 3.6 & 0.1 & 0.3 & -0.5 \\ -0.2 & 0.1 & 8.1 & -0.2 & 1.4 \\ 0.4 & 0.3 & -0.2 & 1.8 & -0.7 \\ 0 & -0.5 & 1.4 & -0.7 & 2.3 \end{bmatrix} \right)$$

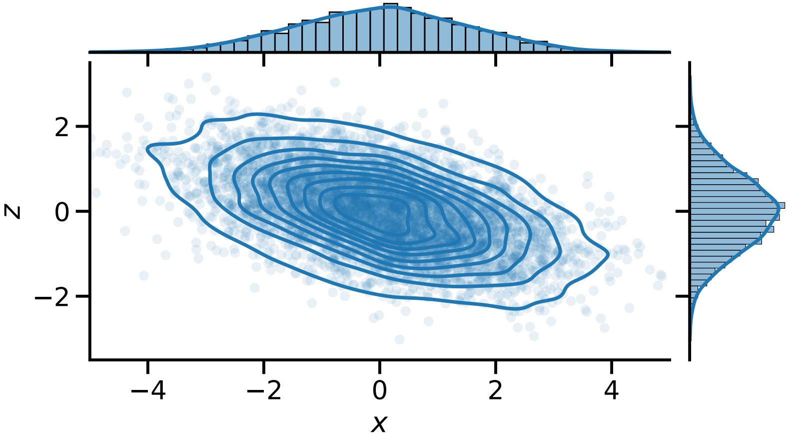
- Notice that $x = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}$, so

$$X \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^\mathsf{T} & \boldsymbol{\Sigma}_z \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}^\mathsf{T} \right)$$

$$X \sim \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x \right)$$

Marginalizing Gaussians

- If $\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^\mathsf{T} & \boldsymbol{\Sigma}_z \end{bmatrix}\right)$, then $x \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$:
  we can just ignore a subset of the variables



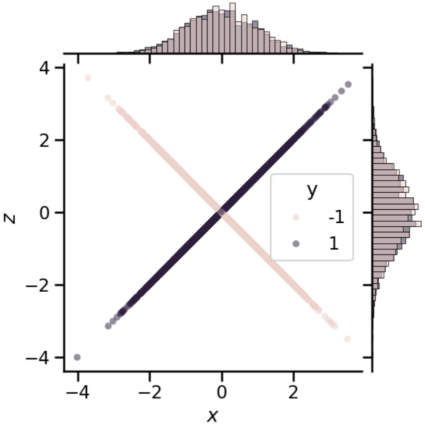
Independence structure in Gaussians

- For bivariate Gaussians, if $\Sigma_{12} = 0$ then Σ is diagonal, and so $x_1 \perp\!\!\!\perp x_2$
- So, in multivariate Gaussians, $x_j \perp\!\!\!\perp x_{j'}$ iff $\Sigma_{jj'} = 0$
- If $\Sigma_{jj'} \neq 0$, x_j and $x_{j'}$ are correlated: can have all pairs correlated
- Multivariate Gaussians don't have any nonlinear or "higher-order" interactions

- Example:

$$x \sim \mathcal{N}(0, 1)$$
$$y \sim \mathrm{Unif}(\{-1, 1\})$$
$$z = xy$$

- $x \perp\!\!\!\perp y$, $\mathrm{Cov}(x, z) = 0$, $y \perp\!\!\!\perp z$
- $x \sim \mathcal{N}(0, 1)$, $z \sim \mathcal{N}(0, 1)$
  - But they're not jointly normal



Conditioning in Gaussians

- If $\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xz} \\ \boldsymbol{\Sigma}_{xz}^\mathsf{T} & \boldsymbol{\Sigma}_z \end{bmatrix} \right)$, then what's $x \mid z$?

- By doing a bunch of linear algebra (see PML1 7.3.5), you get

$$\begin{aligned} x \mid z &\sim \mathcal{N}(\boldsymbol{\mu}_{x|z}, \boldsymbol{\Sigma}_{x|z}) \\ \boldsymbol{\mu}_{x|z} &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xz} \boldsymbol{\Sigma}_z^{-1} (z - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{x|z} &= \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xz} \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\Sigma}_{xz}^\mathsf{T} \end{aligned}$$

- If you know the value of z, the distribution of x is a different Gaussian
- If $\boldsymbol{\sigma}_{xz} = \mathbf{0}$, then $x \mid \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$; another way to see $x \perp\!\!\!\perp z$
- Notice that while $\boldsymbol{\mu}_{x|z}$ depends on the value of z, $\boldsymbol{\Sigma}_{x|z}$ doesn't!
  - This property is occasionally surprisingly important

Outline

MLE for the mean of a multivariate Gaussian

- If $x^{(i)} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma} \succ 0$, we have

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

  so up to a constant our negative log-likelihood for n examples is

$$\frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\left(x^{(i)} - \boldsymbol{\mu}\right) + \frac{n}{2}\log\left|\boldsymbol{\Sigma}\right|$$

- This is a convex quadratic in μ; setting gradient to zero gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

  - Mean along each dimension; it doesn't depend on Σ

MLE for the covariance of a multivariate Gaussian

- To get MLE for Σ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)}-\boldsymbol{\mu}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\left(x^{(i)}-\boldsymbol{\mu}\right)+\frac{n}{2}\log|\boldsymbol{\Sigma}|$$

$$=\frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)}-\boldsymbol{\mu}\right)^{\mathsf{T}}\boldsymbol{\Theta}\left(x^{(i)}-\boldsymbol{\mu}\right)+\frac{n}{2}\log\left|\boldsymbol{\Theta}^{-1}\right|$$

- After some work (bonus slides), we get that this is equal to

$$f(\boldsymbol{\Theta})=\frac{n}{2}\operatorname{Tr}(\mathbf{S}\Theta)-\frac{n}{2}\log|\boldsymbol{\Theta}|,\ \text{with } \mathbf{S}=\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}-\boldsymbol{\mu}\right)\left(x^{(i)}-\boldsymbol{\mu}\right)^{\mathsf{T}}$$

  - $\mathbf{S}$ is the sample covariance: if $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\mu^{\mathsf{T}}$ is centred data, $S = (1/n)\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}}$
  - Trace operator $\operatorname{Tr}(A)$ is the sum of the diagonal elements of A
  - $\operatorname{Tr}(A^{\mathsf{T}}B) = \sum_j (A^{\mathsf{T}}B)_{jj} = \sum_j \sum_i (A^{\mathsf{T}})_{ji}B_{ij} = \sum_{ij} A_{ij}B_{ij}$, i.e. `(A * B).sum()`

MLE for the covariance of a multivariate Gaussian

- Gradient matrix of NLL with respect to Θ is (not obvious, see bonus slides)

$$\nabla f(\Theta) = \frac{n}{2}\left(\mathbf{S} - \mathbf{\Theta}^{-1}\right) \quad \text{for } S = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \boldsymbol{\mu}\right)\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}}$$

- The MLE for a given μ is obtained by setting the gradient matrix to zero, giving

$$\Theta = \mathbf{S}^{-1} \quad \text{or} \quad \Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^{\mathsf{T}}$$

- To have $\Sigma \succ 0$, we need a positive-definite sample covariance, $S \succ 0$
  - If S is not positive definite, NLL is unbounded below, and MLE doesn't exist
  - Like requiring "not all values are the same" in univariate Gaussian
  - In d-dimensions, you need d linearly independent $x^{(i)}$ values (no "multi-collinearity")
  - This is only possible if $n \geq d$! (But might not be true even if it is)

- Note: most distributions' MLEs don't correspond with "moment matching"

Example: Multivariate Gaussians on MNIST

- Let's try continuous density estimation on (binary) handwritten digits

$x^i = \text{vec}\Big($  $\Big) \in \mathbb{R}^{784}$

7 7
4 4

Diagonal Σ:

$\hat{\mu} = \text{vec}\Big($  $\Big)$

$\hat{\Sigma} = \text{diag}\Big(\text{vec}\Big($  $\Big)\Big)$

General Σ:

$\hat{\mu}$ is the same (!)

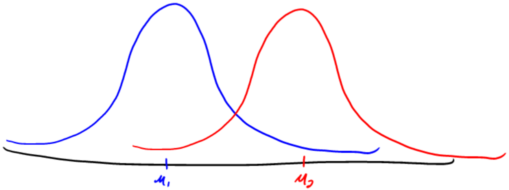
$\hat{\Sigma}$ is big (784 by 784)

Product of Gaussian densities

- This property will be helpful in deriving MAP/Bayesian estimation:
- Consider a variable x whose pdf is written as product of two Gaussians,

$$p(x) \propto \underbrace{\mathcal{N}(x \mid \boldsymbol{\mu}_1, \mathbf{I})}_{\text{density of } \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}) \text{ at } x} \quad \mathcal{N}(x \mid \boldsymbol{\mu}_2, \mathbf{I})$$

- This product of Gaussian pdfs is a Gaussian with $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma} = \frac{1}{2}\mathbf{I}$



Product of Gaussian densities

- If $p(x) \propto \mathcal{N}(x \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\, \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$,
- then x is Gaussian with (see PML2 2.2.7.6 – complete the square in the exponent)

$$\text{covariance} \quad \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$$

$$\text{mean} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

- Consider $x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for fixed $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$:

$$\begin{aligned} p(\boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto p(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} p\left(x^{(i)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) && \text{(Bayes rule)} \\ &= p(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} p(\boldsymbol{\mu} \mid x^{(i)}, \boldsymbol{\Sigma}) && (\text{symmetry of } x^{(i)} \text{ and } \boldsymbol{\mu}) \\ &= \text{(product of } (n+1) \text{ Gaussians)} \end{aligned}$$

- So, working it out gives...

MAP estimation for mean

- For fixed Σ, conjugate prior for mean is a Gaussian:

$$x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{implies} \quad \boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+),$$

  where

$$\boldsymbol{\Sigma}^+ = (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1},$$

$$\boldsymbol{\mu}^+ = \Sigma^+(n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \qquad \text{MAP estimate of } \mu$$

- In special case of $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $\Sigma_0 = \frac{1}{\lambda}\mathbf{I}$, we get

$$\boldsymbol{\Sigma}^+ = \left(\frac{n}{\sigma^2}\mathbf{I} + \lambda\mathbf{I}\right)^{-1} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\lambda}}\mathbf{I},$$

$$\boldsymbol{\mu}^+ = \boldsymbol{\Sigma}^+\left(\frac{n}{\sigma^2}\boldsymbol{\mu}_{\text{MLE}} + \lambda\boldsymbol{\mu}_0\right)$$

- Posterior predictive is $\mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}^+)$ – take product of $(n+2)$ then marginalize
  - Many Bayesian inference tasks have closed form; if not, Monte Carlo is easy

MAP Estimation in Multivariate Gaussian (Trace Regularization)

- A common MAP estimate for Σ is

$$\hat{\Sigma} = \mathbf{S} + \lambda \mathbf{I},$$

  where S is the covariance of the data.
  - Key advantage: $\hat{\Sigma}$ is positive-definite (eigenvalues are at least λ)

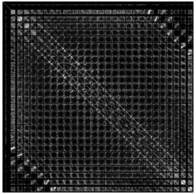
- This corresponds to L1 regularization of precision diagonals (see bonus)

$$f(\Theta) = \underbrace{\mathrm{Tr}(\mathbf{S}\mathbf{\Theta}) - \log|\mathbf{\Theta}|}_{\text{NLL times } 2/n} + \lambda \sum_{j=1}^{d} |\mathbf{\Theta}_{jj}|$$

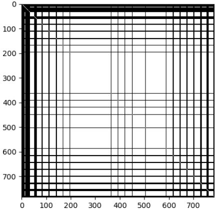
- Note this *doesn't* set Θ_{jj} values to exactly zero
  - Log-determinant term becomes arbitrarily steep as the Θ_{jj} approach 0
  - It's not really the case that "L1 gives sparsity"; it's "L2 + L1 gives sparsity"

Trace Regularization

- For MNIST, MAP estimate of precision $\mathbf{\Theta}$ with regularizer $\frac{1}{n}\operatorname{Tr}(\mathbf{\Theta})$



- Sparsity pattern using this "L1-regularization of the trace":



- Doesn't yield a sparse matrix (only zeroes are with pixels near the boundary)

Summary

- Multivariate Gaussians: random vectors, which allow correlations
- Affine transformations of Gaussians are Gaussian
  - Can use that to sample
- Marginals, conditionals are also Gaussian

MLE for the covariance of a multivariate Gaussian

bonus!

- To get MLE for Σ we re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^\mathsf{T}\Sigma^{-1}(x^i - \mu) + \frac{n}{2}\log|\Sigma|$$

$$= \frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^\mathsf{T}\Theta(x^i - \mu) + \frac{n}{2}\log|\Theta^{-1}| \qquad \text{(okay because } \Sigma \text{ is invertible)}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x^{(i)} - \mu)^\mathsf{T}\Theta(x^i - \mu)\right) + \frac{n}{2}\log|\Theta|^{-1} \qquad \text{(scalar } y^\mathsf{T}Ay = \mathrm{Tr}(y^\mathsf{T}Ay))$$

$$= \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}((x^{(i)} - \mu)(x^i - \mu)^\mathsf{T}\Theta) - \frac{n}{2}\log|\Theta| \qquad (\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB))$$

- $\left|A^{-1}\right| = 1/|A|$ (can see e.g. from eigenvalues)
- The trace is the sum of the diagonal elements: $\mathrm{Tr}(A) = \sum_i A_{ii}$
  - $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ when dimensions match: called trace rotation or cyclic property

MLE for the covariance of a multivariate Gaussian

bonus!

- From the last slide,

$$p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{Tr}\left(\left(x^{(i)} - \boldsymbol{\mu}\right)\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Theta}\right) - \frac{n}{2}\log|\boldsymbol{\Theta}|$$

- We can exchange the sum and trace (trace is a linear operator) to get,

$$= \frac{1}{2}\,\mathrm{Tr}\left(\sum_{i=1}^{n}(x^{(i)} - \mu)(x^{i} - \mu)^{\mathsf{T}}\Theta\right) - \frac{n}{2}\log|\Theta| \qquad \sum_{i}\mathrm{Tr}(A_iB) = \mathrm{Tr}\left(\sum_{i} A_iB\right)$$

$$= \frac{n}{2}\,\mathrm{Tr}\left(\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x^{i} - \mu)(x^{i} - \mu)^{\mathsf{T}}}_{\text{sample covariance, } S}\right)\Theta\right) - \frac{n}{2}\log|\Theta| \qquad \left(\sum_{i} A_iB\right) = \left(\sum_{i} A_i\right)B$$

MLE for the covariance of a multivariate Gaussian

bonus!

- So the NLL in terms of the precision matrix Θ and sample covariance S is

$$f(\Theta) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta|, \text{ with } S = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \boldsymbol{\mu}\right)\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}}$$

- Weird-looking but has nice properties:
  - $\text{Tr}(S\Theta)$ is linear function of Θ, with $\nabla_\Theta \text{Tr}(S\Theta) = S$
    (it's the matrix version of an inner product $s^{\mathsf{T}}\theta$; called "Frobenius inner product")
  - Negative log-determinant is strictly convex, and $\nabla_\Theta \log|\Theta| = \Theta^{-1}$
    (generalizes $\nabla \log|x| = 1/x$ for for $x > 0$)
- Using these two properties the gradient matrix has a simple form:

$$\nabla f(\Theta) = \frac{n}{2}(S - \Theta^{-1})$$

which is what we use to get the MLE