TOOLS: EIGENVALUES AND GAUSSIAN DISTRIBUTIONS

We consider a square matrix $A \in \mathbb{R}^{m \times m}$.

Definition

A vector $\xi \in \mathbb{R}^m$ is called an **eigenvector** of A if the direction of ξ does not change under application of A. In other words, if there is a scalar λ such that

$$A\xi = \lambda\xi \ .$$

λ is called an **eigenvalue** of A for the eigenvector ξ.

Properties in general

- In general, eigenvalues are complex numbers $\lambda \in \mathbb{C}$.
- The class of matrices with the nicest eigen-structure are symmetric matrices, for which all eigenvalues are real numbers.

EIGENSTRUCTURE OF SYMMETRIC MATRICES

If a matrix is symmetric:

- All eigenvalues and eigenvectors are real, i.e. $\lambda \in \mathbb{R}$ and $\xi \in \mathbb{R}^m$.
- There are $\text{rank}(A)$ distinct eigenvectors.
- The eigenvectors are pair-wise orthogonal.
- If $\text{rank}(A) = m$, there is an ONB of $\mathbb{R}^m$ consisting of eigenvectors of A.

Definiteness

type	if ...
positive definite	all eigenvalues > 0
positive semi-definite	all eigenvalues ≥ 0
negative semi-definite	all eigenvalues ≤ 0
negative definite	all eigenvalues < 0
indefinite	none of the above

Eigenvector ONB

Setting

- Suppose A symmetric, $\xi_1, \dots, \xi_m$ are eigenvectors and form an ONB.
- $\lambda_1, \dots, \lambda_m$ are the corresponding eigenvalues.

How does A act on a vector $v \in \mathbb{R}^m$?

1. Represent v in basis $\xi_1, \dots, \xi_m$:

$$v = \sum_{j=1}^{m} v_j^{\mathrm{A}} \xi_j \qquad \text{where } v_j^{\mathrm{A}} \in \mathbb{R}$$

2. Multiply by A: Eigenvector definition (recall: $A\xi_j = \lambda\xi_j$) yields

$$Av = A\Big(\sum_{j=1}^{m} v_j^{\mathrm{A}} \xi_j\Big) = \sum_{j=1}^{m} v_j^{\mathrm{A}} A\xi_j = \sum_{j=1}^{m} v_j^{\mathrm{A}} \lambda_j \xi_j$$

Conclusion

A symmetric matrix acts by scaling the directions ξ_j.

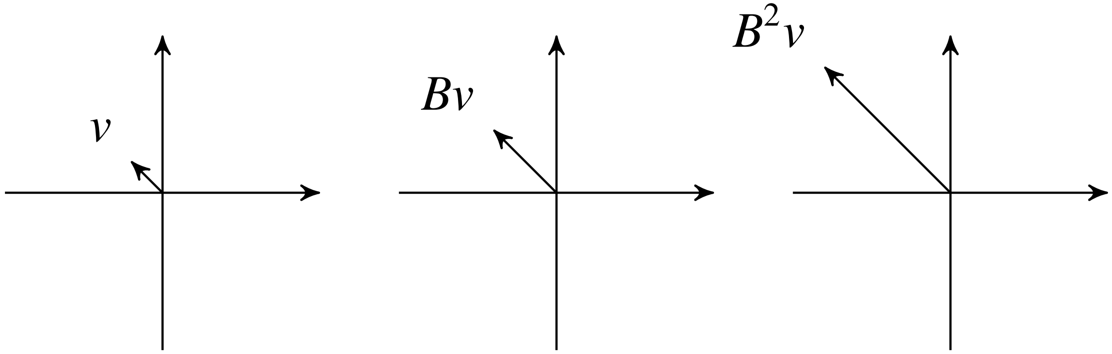
Setting

We *repeatedly* apply a symmetric matrix B to some vector $v \in \mathbb{R}^m$, i.e. we compute

$$Bv, \quad B(Bv) = B^2 v, \quad B(B(Bv))) = B^3 v, \quad \ldots$$

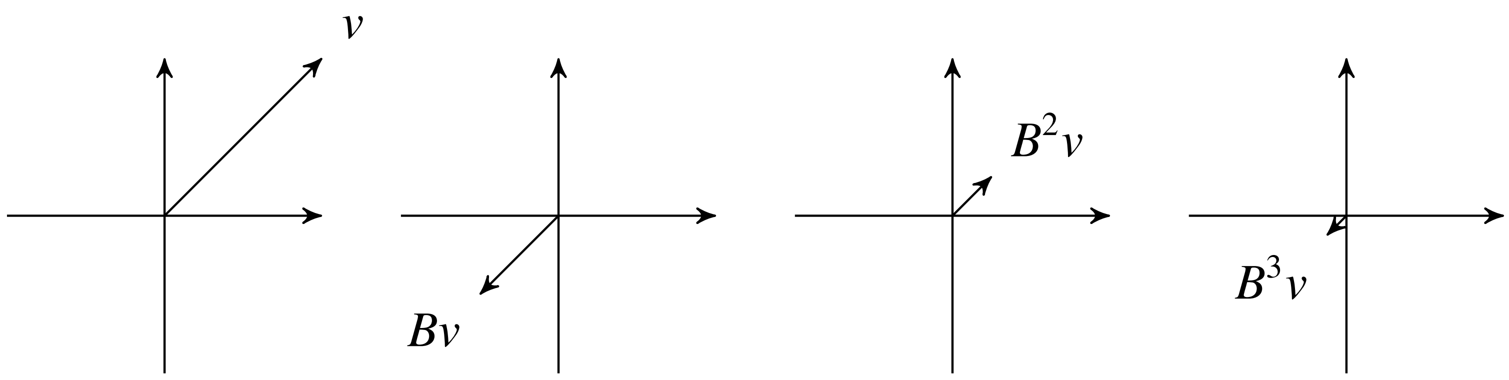
How does v change?

Example 1: v is an eigenvector with eigenvalue 2



The direction of v does not change, but its length doubles with each application of B.

Example 2: v is an eigenvector with eigenvalue $-\frac{1}{2}$



For an arbitrary vector v

$$B^n v = \sum_{j=1}^{m} v_j^{\mathrm{B}} \lambda_j^n \xi_j$$

- The weight λ_j^n grows most rapidly for eigenvalue with largest absolute value.
- Consequence:

The direction of $B^n v$ converges to the direction of the eigenvector with largest eigenvalue as n grows large.

QUADRATIC FORMS

In applications, symmetric matrices often occur in quadratic forms.

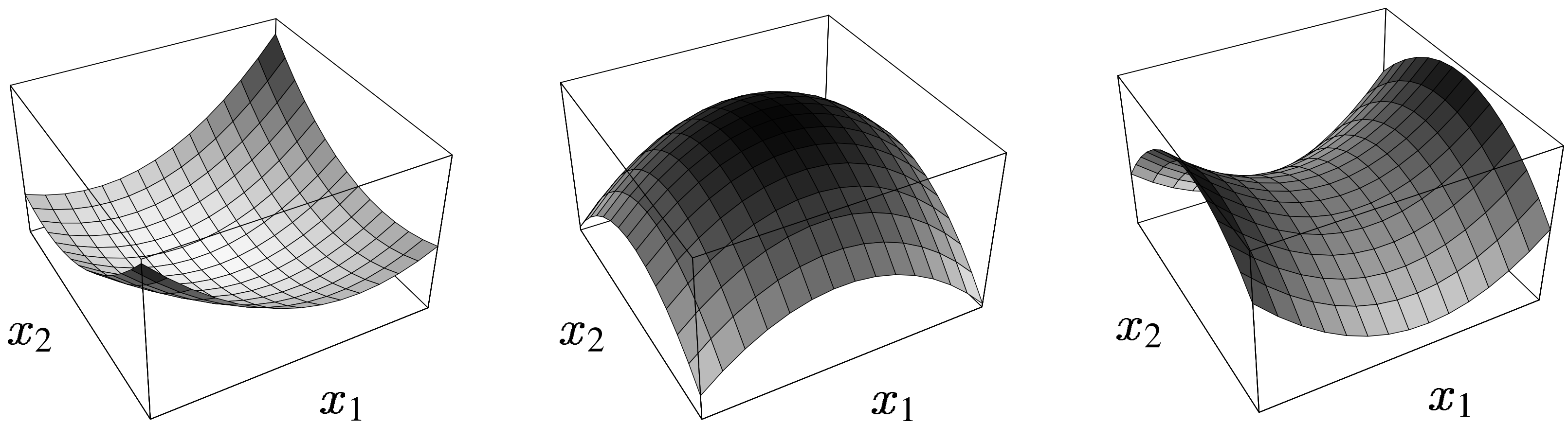
Definition

The **quadratic form** defined by a matrix A is the function

$$
\begin{aligned}
q_A :\ \mathbb{R}^m &\to \mathbb{R} \\
x &\mapsto \langle x, Ax \rangle
\end{aligned}
$$

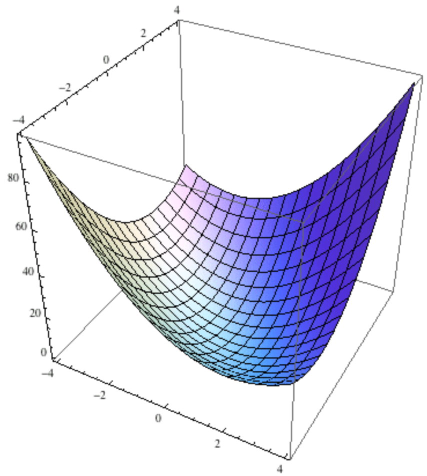
Intuition

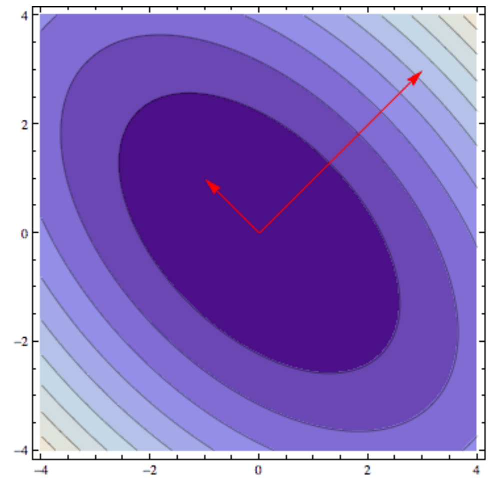
A quadratic form is the m-dimensional analogue of a quadratic function ax^2, with a vector substituted for the scalar x and the matrix A substituted for the scalar $a \in \mathbb{R}$.



Quadratic Forms

Here is the quadratic form for the matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$:

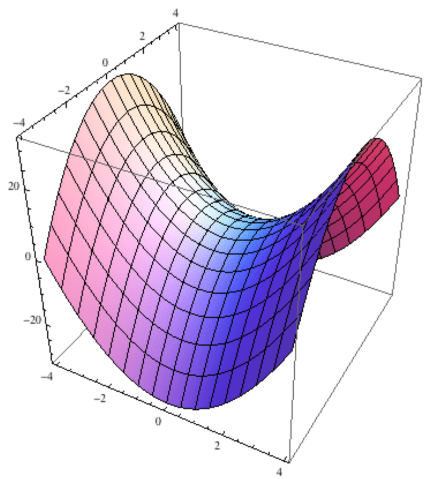


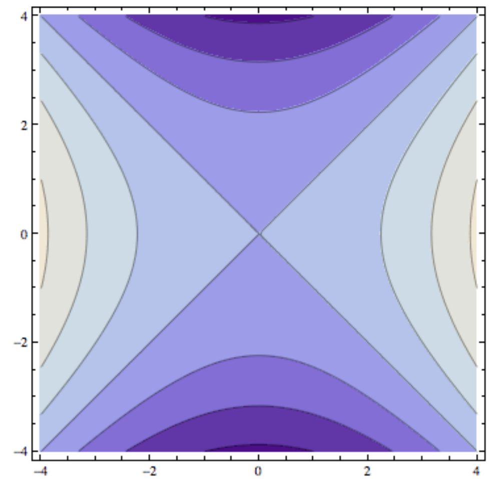


- Left: The function value q_A is graphed on the vertical axis.
- Right: Each line in $\mathbb{R}^2$ corresponds to a constant function value of q_A.
  Dark color = small values.
- The red lines are eigenvector directions of A. Their lengths represent the (absolute) values of the eigenvalues.
- In this case, both eigenvalues are positive. If all eigenvalues are positive, the contours are ellipses. So:

$$\text{positive definite matrices} \quad \leftrightarrow \quad \text{elliptic quadratic forms}$$

In this plot, the eigenvectors are axis-parallel, and one eigenvalue is negative:





The matrix here is $A = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$.

Intuition

- If we change the sign of one of the eigenvalue, the quadratic function along the corresponding eigen-axis flips.
- There is a point which is a minimum of the function along one axis direction, and a maximum along the other. Such a point is called a *saddle point*.

Recall: Covariance

The covariance of two random variables X_1, X_2 is

$$\mathrm{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] \; .$$

If $X_1 = X_2$, the covariance is the variance: $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$.

Covariance matrix

If $X = (X_1, \ldots, X_m)$ is a random vector with values in $\mathbb{R}^m$, the matrix of all covariances

$$\mathrm{Cov}[X] := (\mathrm{Cov}[X_i, X_j])_{i,j} = \begin{pmatrix} \mathrm{Cov}[X_1, X_1] & \cdots & \mathrm{Cov}[X_1, X_m] \\ \vdots & & \vdots \\ \mathrm{Cov}[X_m, X_1] & \cdots & \mathrm{Cov}[X_m, X_m] \end{pmatrix}$$

is called the **covariance matrix** of X.

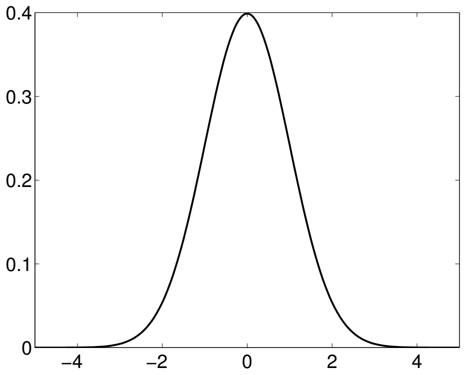
Notation

It is customary to denote the covariance matrix $\mathrm{Cov}[X]$ by Σ.

GAUSSIAN DISTRIBUTION

Gaussian density in one dimension

$$p(x;\mu,\sigma) := \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

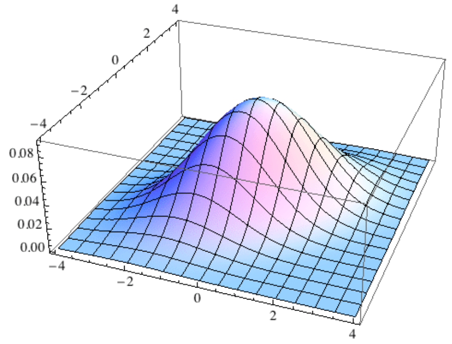


- μ = expected value of x, σ^2 = variance, σ = standard deviation
- The quotient $\frac{x-\mu}{\sigma}$ measures deviation of x from its expected value in units of σ (i.e. σ defines the length scale)

Gaussian density in m dimensions

The quadratric function

$$-\frac{(x-\mu)^2}{2\sigma^2} \quad = \quad -\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)$$

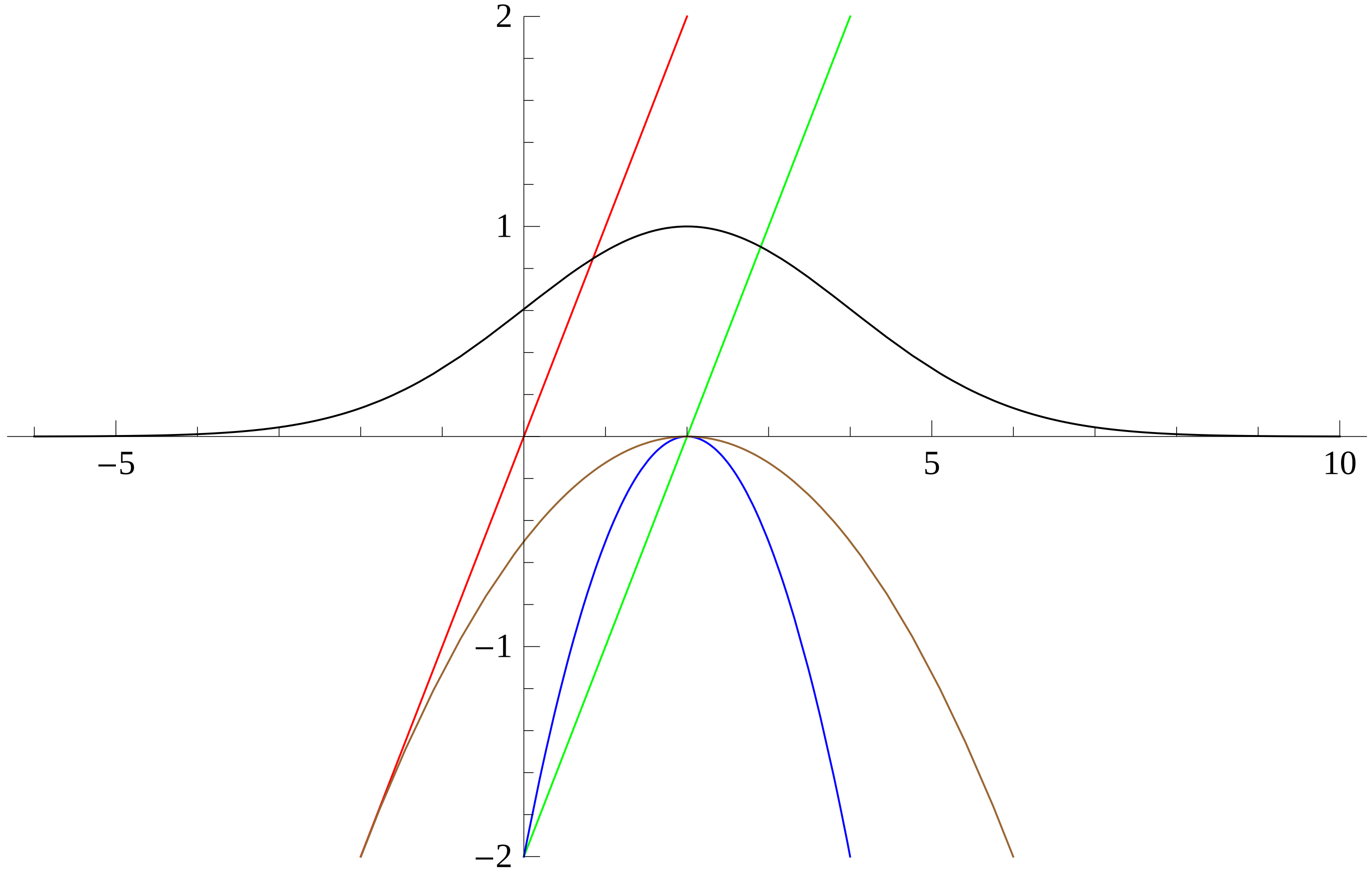


is replaced by a quadratic form:

$$p(\mathbf{x};\boldsymbol{\mu},\Sigma) := \frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\left(-\frac{1}{2}\left\langle(\mathbf{x}-\boldsymbol{\mu}),\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\rangle\right)$$

Components of a 1D Gaussian

$\mu = 2, \sigma = 2$



- Red: $x \mapsto x$
- Green: $x \mapsto x - \mu$
- Blue: $x \mapsto -\frac{1}{2}(x - \mu)^2$
- Brown: $x \mapsto -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$
- Black: $x \mapsto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

Covariance matrix of a Gaussian

If a random vector $X \in \mathbb{R}^m$ has Gaussian distribution with density $p(\mathbf{x}; \mu, \Sigma)$, its covariance matrix is $\text{Cov}[X] = \Sigma$. In other words, a Gaussian is parameterized by its covariance.

Observation

Since $\text{Cov}[X_i, X_j] = \text{Cov}[X_j, X_i]$, the covariance matrix is symmetric.

What is the eigenstructure of Σ?

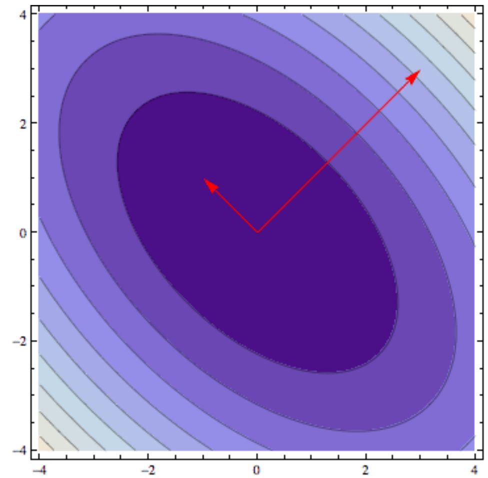
- We know: Σ symmetric $\Rightarrow$ there is an eigenvector ONB
- Call the eigenvectors in this ONB $\xi_1, \dots, \xi_m$ and their eigenvalues $\lambda_1, \dots, \lambda_m$
- We can rotate the coordinate system to $\xi_1, \dots, \xi_m$. In the new coordinate system, Σ has the form

$$\Sigma_{[\xi_1,\dots,\xi_n]} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix} = \text{diag}(\lambda_1, \dots, \lambda_m)$$

Quadratic form

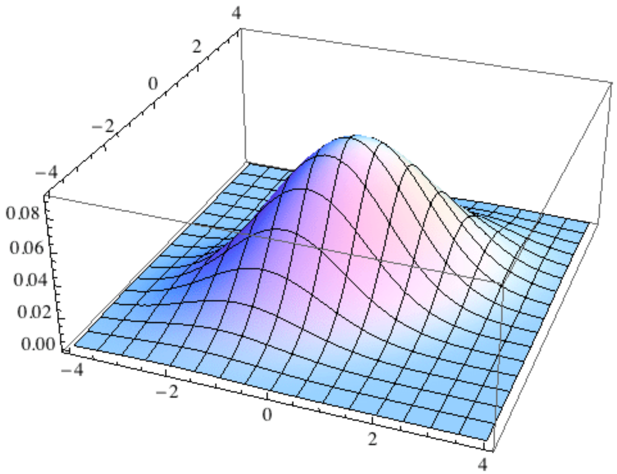
$$\langle \mathbf{x}, \Sigma \mathbf{x} \rangle \quad \text{with} \quad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

The eigenvectors are $(1, 1)$ and $(-1, 1)$ with eigenvalues 3 and 1.

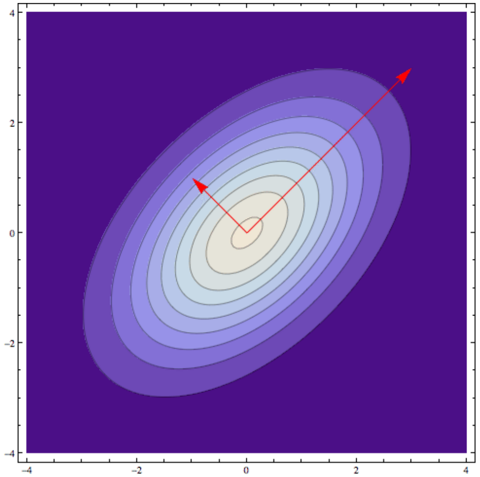


Gaussian density

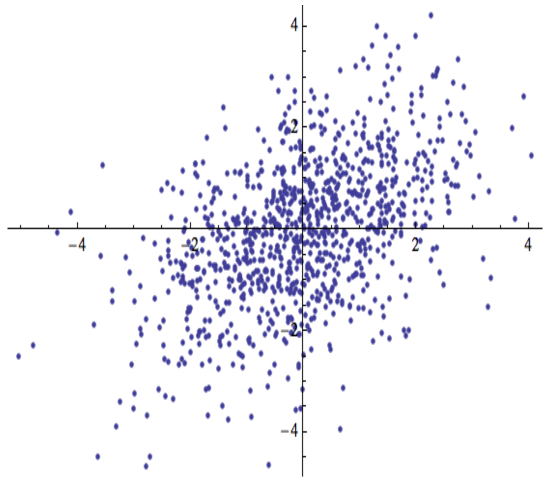
$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0, 0)$.



Density graph



Density contour



1000 sample points

The ξ_i as random variables

Write $e_1, \ldots, e_m$ for the ONB of axis vectors. We can represent each ξ_i as

$$\xi_i = \sum_{j=1}^{m} \alpha_{ij} e_j$$

Then $O = (\alpha_{ij})$ is the orthogonal transformation matrix between the two bases. We can represent random vector $X \in \mathbb{R}^m$ sampled from the Gaussian in the eigen-ONB as

$$X_{[\xi_1,\ldots,\xi_m]} = (X'_1, \ldots, X'_m) \qquad \text{with} \qquad X'_i = \sum_{j=1}^{m} \alpha_{ij} X_j$$

Since the X_j are random variables (and the α_{ij} are fixed), each X'_i is a scalar random variable.

Meaning of the random variables ξ_i

For any Gaussian $p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$, we can

1. shift the origin of the coordinate system into $\boldsymbol{\mu}$
2. rotate the coordinate system to the eigen-ONB of Σ.

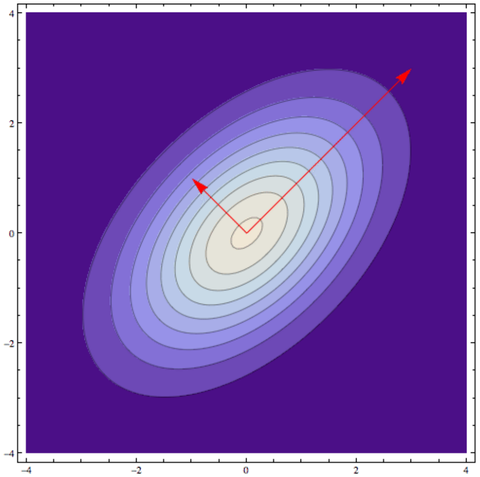
In this new coordinate system, the Gaussian has covariance matrix

$$\Sigma_{[\xi_1,\ldots,\xi_m]} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$$

where λ_i are the eigenvalues of Σ.

Gaussian in the new coordinates

A Gaussian vector $X_{[\xi_1,\ldots,\xi_m]}$ represented in the new coordinates consists of m *independent* 1D Gaussian variables X_i'. Each X_i' has mean 0 and variance λ_i.



SHRINKAGE

Issues with Least Squares

Robustness

- Least squares works only if $\tilde{\mathbf{X}}$ has full column rank, i.e. if $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ is invertible.
- If $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ *almost* not invertible, least squares is numerically unstable.
  Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

- Modern problems: Many dimensions/features/predictors (possibly thousands)
- Only a few of these may be important
  $\rightarrow$ need some form of feature selection
- Least squares:
  - Treats all dimensions equally
  - Relevant dimensions are averaged with irrelevant ones
  - Consequence: Signal loss

Regularity of Matrices

Regularity

A matrix which is not invertible is also called a **singular** matrix. A matrix which is invertible (not singular) is called **regular**.

In computations

Numerically, matrices can be "almost singular". Intuition:

- A singular matrix maps an entire linear subspace into a single point.
- If a matrix maps points far away from each other to points very close to each other, it almost behaves like a singular matrix.

Regularity of Symmetric Matrices

Recall: A positive semi-definite matrix A is singluar $\Leftrightarrow$ smallest EValue is 0

Illustration

If smallest EValue $\lambda_{\min} > 0$ but very small (say $\lambda_{\min} \approx 10^{-10}$):

- Suppose x_1, x_2 are two points in subspace spanned by $\xi_{\min}$ with $\|x_1 - x_2\| \approx 1000$.
- Image under A: $\|Ax_1 - Ax_2\| \approx 10^{-7}$

In this case

- A has an inverse, but A behaves almost like a singular matrix
- The inverse A^{-1} can map almost identical points to points with large distance, i.e.

  small change in input $\quad\rightarrow\quad$ large change in output

  $\rightarrow$ unstable behavior

Consequence for Statistics

If a statistical prediction involves the inverse of an almost-singular matrix, the predictions become unreliable (high variance).

Implications for Linear Regression

Recall: Prediction in linear regression

For a point $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$, we predict the corresponding function value as

$$\hat{y}_{\text{new}} = \left\langle \hat{\beta}, (1, \mathbf{x}) \right\rangle = (\tilde{\mathbf{X}}^t \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^t \mathbf{y}$$

Effect of unstable inversion

- Suppose we choose an arbitrary training point $\tilde{\mathbf{x}}_i$ and make a small change to its response value $\tilde{y}_i$.
- Intuitively, that should not have a big impact on $\hat{\beta}$ or on prediction.
- If $\tilde{\mathbf{X}}^t \tilde{\mathbf{X}}$ is almost singular, a small change to $\tilde{y}_i$ can prompt a huge change in $\hat{\beta}$, and hence in the predicted value $\hat{y}_{\text{new}}$.