Mixtures of Linear-Gaussian Models

Mathias Winther Madsen

January 30, 2017

This document spells out the relationship between linear-Gaussian models and jointly Gaussian models in the input-output space. This correspondence is used to make some observations about how to approximate a mixture of two linear-Gaussian models with another linear-Gaussian model.

Contents

Notation:

$.$	matrix multiplication	^{T}	matrix transposition
det	determinant	E	expected value

1 Preliminaries

This section rehearses a few concepts from statistics that will be used later.

1.1 Covariance Matrices

In the one-dimensional case, the variability of a random variable can be described in terms of its **variance**,

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2] - E[X]^2. \end{aligned}$$

The variance can be generalized to the one-dimensional **covariance**,

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y], \end{aligned}$$

which measures the strength of the linear relationship between X and Y.

The corresponding statistic in higher dimensions is the **covariance matrix**,

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X]).(Y - E[Y])^T] \\ & E[X.Y^T] - E[X].E[Y]^T. \end{aligned}$$

The entries of this matrix are the one-dimensional covariance statistics

$$Cov[X, Y]_{ij} = Cov[X_i, Y_j].$$

When $X = Y$, the diagonal terms of the covariance matrix contains the marginal variances of the random coordinates $X_1, X_2, \ldots, X_n$. The off-diagonal terms record the covariance between two distinct entries $X_i \neq X_j$.

The covariance function satisfies the following rules:

$$\begin{aligned} Cov[X, Y] &= Cov[Y, X]^T \\ Cov[X + Y, Z] &= Cov[X, Z] + Cov[Y, Z] \\ Cov[Z, X + Y] &= Cov[Z, X] + Cov[Z, Y] \\ Cov[A.X, Y] &= A.Cov[X, Y] \\ Cov[Y, A.X] &= Cov[X, Y].A^T \end{aligned}$$

For independent X and Y, we moreover have $Cov[X, Y] = \mathbf{0}$ and $Cov[Y, X] = \mathbf{0}$ (but note that the dimensionality of these two zero matrices may be different). Since constants are independent of all other random variables, we also have $Cov[X, c] = 0$ for any deterministic variable c.

1.2 Multivariate Gaussians

Suppose that X is a vector whose coordinates are independent one-dimensional standard Gaussians,

$$X_i \sim \mathcal{N}(0,1) \qquad i = 1, 2, 3, \ldots, n.$$

Then the random vector X is a **standard multivariate Gaussian**.

Since the coordinates of X are assumed independent of each other,

$$E[X_i X_j] \;=\; \begin{cases} 0 & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$

Hence, the the covariance matrix of X is the identity matrix. Its mean is the zero vector.

Now suppose that A is some non-singular matrix and B a vector of the same dimensionality as X. Then the random variable

$$Y \;=\; A.X + B$$

is called a (non-standard) **multivariate Gaussian**. Figure 1 shows an example of such a transformed standard Gaussian with $B = \mathbf{0}$.

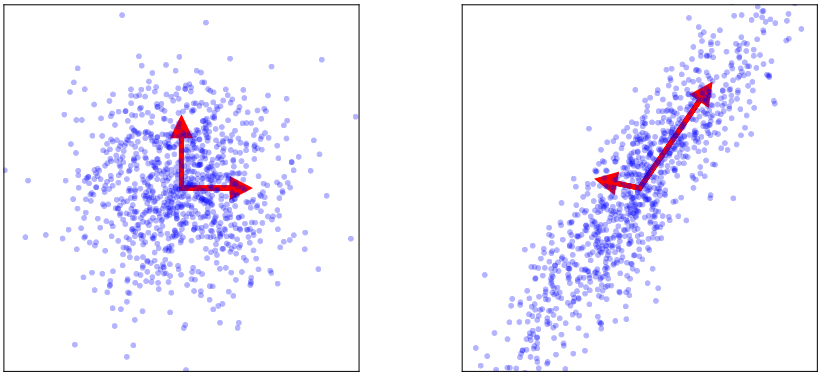

Figure 1: A sample from a standard multivariate Gaussian X, and its deformed cousin $A.X$. Note that the distribution of $A.X$ exhibits both mirror symmetries (invariance with respect to sign flipping) and axis permutation symmetries (invariance with respect to the source of randomness).

By the linearity of expectations, we have $E[Y] = B$, while

$$\begin{aligned} Cov[Y,Y] &= E[(Y - E[Y]).(Y - E]Y])^T] \\ &= E[(A.X).(A.X)^T] \\ &= A.E[X.X^T].A^T \\ &= A.A^T, \end{aligned}$$

since X was assumed to be a standard multivariate Gaussian.

The multivariate Gaussian is usually parametrized in terms of

$$\begin{aligned} \mu &= B \\ \Sigma^2 &= A.A^T \end{aligned}$$

If we want to recover the matrix A, we therefore have to find its "square root" by means of a matrix decomposition technique such as Eigenvalue decomposition (or singular value decomposition, which here amounts to the same thing).

2 Linear-Gaussian and Jointly Gaussian Models

Suppose that X is an n-dimensional Gaussian, and that A is an $m \times n$ matrix. Then $Y = A.X$ is an m-dimensional Gaussian, and the stacked vector

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix},$$

which below is abbreviated as $Z = (X, Y)$, is an $(n+m)$-dimensional Gaussian.

This defines a correspondence between **linear-Gaussian models** (that is, linear functions $f : \mathbb{R}^n \to \mathbb{R}^m$ with Gaussian inputs) and **jointly Gaussian distributions** over $\mathbb{R}^{n+m}$, as shown in Figure 2. The following subsections collect some formulas of relevance to this correspondence.

2.1 From Jointly Gaussian to Linear-Gaussian Models

Suppose the pair $Z = (X, Y)$ is a Gaussian vector with a mean of

$$E[Z] = \begin{pmatrix} E[X] \\ E[Y] \end{pmatrix}$$

and a covariance given by the block matrix

$$Cov[Z] = \begin{pmatrix} Cov[X,X] & Cov[X,Y] \\ Cov[Y,X] & Cov[Y,Y] \end{pmatrix}.$$

Then Y can equivalently be defined as a noisy, linear transformation of X:

$$Y = A.X + B + \varepsilon.$$

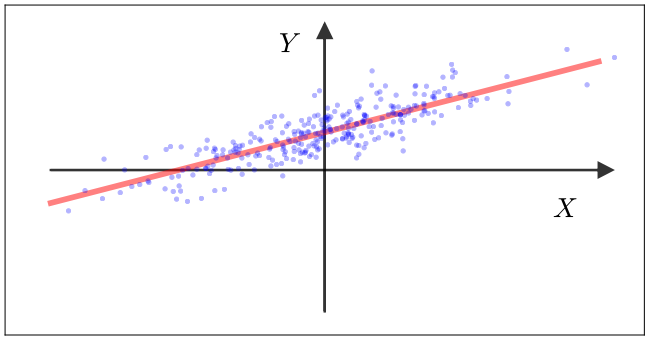

Figure 2: The correspondence between linear functions (the red line) and jointly Gaussian distributions (the blue data points).

The slope matrix and intercept vector of this linear transformation are then

$$\begin{aligned} A &= Cov[Y,X].Cov[X,X]^{-1} \\ &= Cov[X,X]^{-1}.Cov[X,Y] \\ B &= E[Y] - A.E[X] \end{aligned}$$

and the noise variable ε is a Gaussian with

$$\begin{aligned} E[\varepsilon] &= 0 \\ Cov[\varepsilon,\varepsilon] &= Cov[Y,Y] - A.Cov[X,X].A^T. \end{aligned}$$

Note the following special cases:

- When X is a normalized Gaussian, we have $A = Cov[Y,X]$. In all other cases, we need to compensate for the scale-dependence of $Cov[Y,X]$ by right-multiplying by the normalizing factor $Cov[X,X]^{-1}$.
- When $\varepsilon = 0$, we have $Cov[Y,Y] = A.Cov[X,X].A^T$. The variation in Y is thus explained completely in terms of the variation in X. In the presence of non-trivial noise, the remaining variation in Y is attributed to $Cov[\varepsilon,\varepsilon]$.

2.2 From Linear-Gaussian to Jointly Gaussian Models

Suppose that the random variable Y is defined by

$$Y = A.X + B + \varepsilon,$$

where ε is a Gaussian noise variable with $E[\varepsilon] = 0$ and $Cov[\varepsilon] = \Phi^2$.

This formulas defines a conditional distribution for Y given X, but not marginal distribution over X. In order to derive a joint distribution for the pair (X, Y), we therefore need to assume a marginal distribution over X.

Suppose therefore that X is Gaussian distribution random vector with

$$\begin{aligned} E[X] &= \mu \\ Cov[X] &= \Sigma^2 \end{aligned}$$

Then $Z = (X, Y)$ is also jointly Gaussian, with

$$E[Z] = \begin{pmatrix} \mu \\ A.\mu + B \end{pmatrix}$$

and

$$Cov[Z] = \begin{pmatrix} \Sigma^2 & \Sigma^2.A^T \\ A.\Sigma^2 & A.\Sigma^2.A^T + \Phi^2 \end{pmatrix}.$$

This distribution has the following notable limit cases:

- When $\Sigma^2 \to \mathbf{0}$, we have

$$Cov[Z] \to \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Phi^2 \end{pmatrix}.$$

  This is a singular matrix, reflecting the fact that (X, Y) is concentrated on the infinitely thin slice $\{(y, \mu) : y \in \mathbb{R}\}$ when the input value X is chosen deterministically.

- When $\Phi^2 \to 0$, we have

$$Cov[Z] \to \begin{pmatrix} \Sigma^2 & \Sigma^2.A^T \\ A.\Sigma^2 & A.\Sigma^2.A^T \end{pmatrix}.$$

  By the Dieudonné rule, the determinant of this matrix is

$$\det \left|\Sigma^2\right| \det \left|(A.\Sigma^2.A^T) - (A.\Sigma^2).(\Sigma^2)^{-1}.(\Sigma^2.A^T)\right| = 0.$$

  This reflects the fact that, in the absence of noise, (X, Y) is concentrated on the infinitely thin line $\{(x, A.x + B) \,|\, x \in \mathbb{R}\}$.

The special cases are illustrated in Figure 3.

2.3 From Linear-Gaussian to Linear-Gaussian Models

As the formulas above indicate, the relationship between jointly Gaussian distributions and linear-Gaussian models is informally as follows:

$$\text{Jointly Gaussian} = \text{Gaussian input} + \text{Linear Function} + \text{Gaussian noise}.$$

A jointly Gaussian distribution over (X, Y) thus contains strictly more information than the functional relationship $Y = A.X + B + \varepsilon$. More precisely:

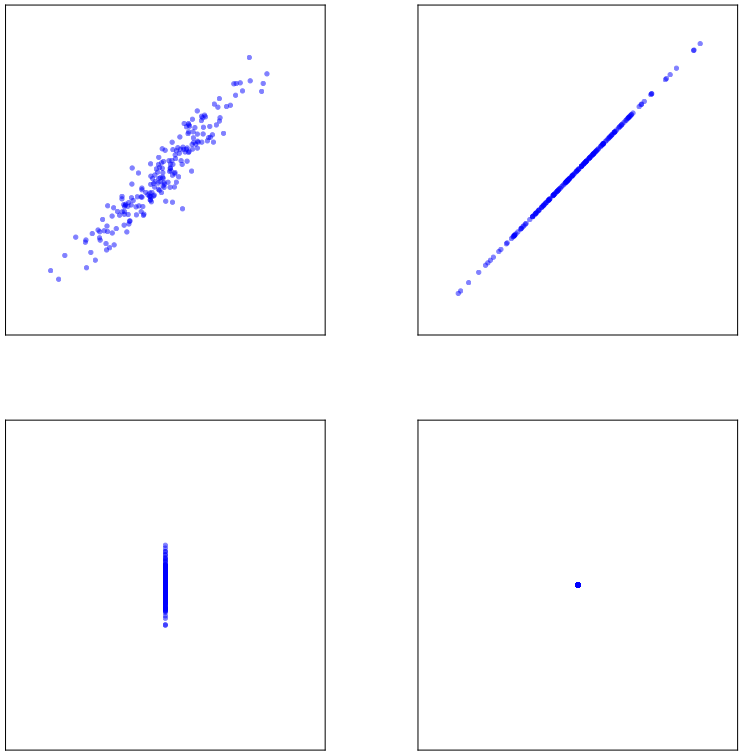

Figure 3: Degenerate linear-Gaussian models: vanishing output noise (right column), and vanishing input variation (bottom row).

- a jointly Gaussian distribution over (X, Y) uniquely determines the functional parameters A and B as well as the marginal distributions over X and ε;
- a functional relationship $Y = A.X + B + \varepsilon$ defines a class of joint distributions over (X, Y), but only when X and ε are Gaussian will (X, Y) be Gaussian too.

Given this asymmetry, it might be worth noticing that when we derive a jointly Gaussian distribution from a linear relationship, and then derive a linear relationship from that joint Gaussian, we get the original linear relationship back.

To see this, let the functional parameters A and B given given. We then choose a Gaussian input distribution $X \sim \mathcal{N}(\mu, \Sigma^2)$ and a Gaussian noise dis-

tribution $\varepsilon \sim \mathcal{N}(0, \Phi^2)$, leading to the statistics

$$\begin{aligned} E[Y] &= A.\mu + B \\ Cov[Y, X] &= A.\Sigma^2 \end{aligned}$$

From these statistics, we obtain the linear parameters

$$\hat{A} \;=\; Cov[Y, X].Cov[X, X]^{-1} \;=\; (A.\Sigma^2).(\Sigma^2)^{-1} \;=\; A;$$

$$\hat{B} \;=\; E[Y] - A.E[X] \;=\; (A.\mu + B) - A.\mu \;=\; B.$$

Since $(\hat{A}, \hat{B}) = (A, B)$, we conclude that the introduction of spurious distributions for X and ε has no effect on the inferred linear relationship between X and Y, whose slope and intercept does not depend on these choices.

3 Gaussian Approximations to Mixture Models

The multivariate Gaussian distribution can be parametrized in terms of its mean vector and its covariance matrix.

As in the univariate case, the maximum likelihood Gaussian fit to a given data set is given by the Gaussian whose mean and covariance that matches the empirical mean and covariance of the data set. Similarly, the best Gaussian approximation to a distribution P is the Gaussian that matches the mean and covariance of P.

In this section, these facts are applied to the case in which we want to approximate a mixture of two probability distributions. The following suctions thus collect some formulas for the mean and covariance of such mixtures, with a special emphasis on the case of mixture components that are derived from linear-Gaussian models.

3.1 Mixture Distributions

Suppose that X_1 and X_2 are two random variables with

$$\begin{aligned} E[X_i] &= \mu_i \\ Cov[X_i] &= \Sigma_i^2 \end{aligned}$$

and that the numbers $p_1, p_2 \in [0, 1]$ satisfy the convexity condition $p_1 + p_2 = 1$. Then the two variables X_1 and X_2 follow certain distributions P_1 and P_2, and these distributions define a mixture $P = p_1 P_1 + p_2 P_2$.

This distribution can also be defined in terms of a sampling program. Set $G_1 \sim \text{Bernoulli}(p_1)$ and, for convenience, $G_2 = 1 - G_1$. Define

$$Z \;=\; G_1 X_1 + G_2 X_2.$$

Then the distribution of Z is the mixture of the distributions of X_1 and X_2. For instance, when $X_1 \sim N(\mu_1, \Sigma_1^2)$ and $X_2 \sim N(\mu_2, \Sigma_2^2)$, Z follows a bimodal

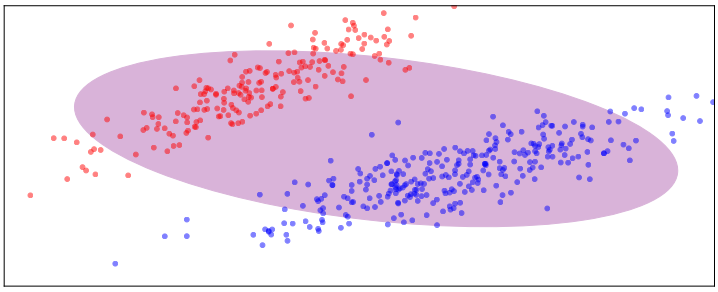

Figure 4: A sample from a mixture of two Gaussians, showing the components in red and blue, and the best Gaussian fit as a purple ellipse.

distribution with peaks at μ_1 and μ_2. (The random average $p_1 X_1 + p_2 X_2$, by contrast, is just a Gaussian random variable with a single peak at $p_1\mu_1 + p_2\mu_2$.)

By the arguments provided in next subsection, we have

$$\begin{aligned} E[Z] &= p_1\mu_1 + p_2\mu_2 \\ Cov[Z,Z] &= p_1\Sigma_1^2 + p_2\Sigma_2^2 + p_1 p_2 (\mu_1 - \mu_2).(\mu_1 - \mu_2)^T. \end{aligned}$$

The means of two mixture components thus combine in a straightforward way, while the covariance has to be adjusted upwards by a term reflecting the distance between the two two mixture components, and their similarity in size.

As mentioned above, this mean vector and covariance matrix also identify the best Gaussian approximation to the mixture distribution, as illustrated in Figure 4. In an online estimation setting, they can also be used to update the mean and covariance statistics of an existing data set given a new batch of data.

3.2 Mean and Covariance of Mixture Distributions

Let $Z = G_1 X_1 + G_2 X_2$ be the mixture distribution described above. Using the independence of G_1 and thus $G_2 = 1 - G_1$ from X_1 and X_2, we then find that

$$\begin{aligned} E[Z] &= E[G_1 X_1 + G_2 X_2] && \text{(definition)} \\ &= E[G_1 X_1] + E[G_2 X_2] && \text{(linearity)} \\ &= E[G_1]E[X_1] + E[G_2]E[X_2] && \text{(indep.)} \\ &= p_1\mu_1 + p_2\mu_2. && \text{(definition)} \end{aligned}$$

In order to find the covariance of Z, we first note that

$$E[G_1] \;=\; E[G_1^2] \;=\; p_1,$$

$$E[G_2] \;=\; E[G_2^2] \;=\; p_2,$$

$$E[G_1 G_2] \;=\; E[G_1(1-G_1)] \;=\; 0,$$

since G_1 and G_2 are binary. Since they are also independent of X_1 and X_2,

$$\begin{aligned} E[Z.Z^T] &= E[(G_1X_1 + G_2X_2).(G_1X_1 + G_2X_2)^T] \\ &= E[G_1^2]\,E[X_1.X_1^T] \;+\; E[G_2^2]\,E[X_2.X_2^T] + \\ &\qquad E[G_1G_2]\,E[X_1.X_2^T] \;+\; E[G_2G_1]\,E[X_2.X_1^T] \\ &= p_1\,E[X_1.X_1^T] \;+\; p_2\,E[X_2.X_2^T]. \end{aligned}$$

Moreover,

$$\begin{aligned} E[Z].E[Z]^T &= (p_1\mu_1 \;+\; p_2\mu_2).(p_1\mu_1 \;+\; p_2\mu_2)^T \\ &= \underbrace{p_1^2(\mu_1.\mu_1^T) + p_2^2(\mu_2.\mu_2^T)}_{\text{squares}} \;+\; \underbrace{p_1p_2(\mu_1.\mu_2^T + \mu_2.\mu_1^T)}_{\text{cross-terms}} \\ &= \underbrace{p_1(1-p_2)(\mu_1.\mu_1^T) + p_2(1-p_1)(\mu_2.\mu_2^T)}_{\text{squares}} \;+\; \underbrace{p_1p_2(\mu_1.\mu_2^T + \mu_2.\mu_1^T)}_{\text{cross-terms}} \\ &= \underbrace{p_1(\mu_1.\mu_1^T) + p_2(\mu_2.\mu_2^T)}_{\text{half the squares}} \;-\; \underbrace{p_1p_2(\mu_1-\mu_2).(\mu_1-\mu_2)^T}_{\text{other half + cross-terms}}. \end{aligned}$$

We put these results together using the fact that

$$\begin{aligned} Cov[Z,Z] &= E[Z.Z^T] \;-\; E[Z].E[Z]^T \\ \Sigma_1^2 &= p_1E[X_1.X_1^T] \;-\; p_1(\mu_1.\mu_1^T) \\ \Sigma_2^2 &= p_2E[X_2.X_2^T] \;-\; p_2(\mu_2.\mu_2^T) \end{aligned}$$

to conclude that

$$Cov[Z,Z] \;=\; p_1\Sigma_1^2 \;+\; p_2\Sigma_2^2 \;+\; p_1p_2(\mu_1-\mu_2).(\mu_1-\mu_2)^T.$$

This confirms the formulas stated in the previous subsection.

3.3 Mixtures of Linear-Gaussian Models

As a special case of the situation above, suppose that the random pairs $Z_i = (X_i, Y_i)$ are defined in terms of the linear-Gaussian models

$$\begin{aligned} X_i &\sim \mathcal{N}(\mu_i, \Sigma_i^2) \\ \varepsilon_i &\sim \mathcal{N}(\mathbf{0}, \Phi_i^2) \\ Y_i &= A_i.X_i + B_i + \varepsilon_i \end{aligned}$$

for $i = 1, 2$. We then have the moments

$$\begin{aligned} E[Z_i] &= \begin{pmatrix} \mu_i \\ A_i.\mu_i + B_i \end{pmatrix} \\ Cov[Z_i] &= \begin{pmatrix} \Sigma_i^2 & \Sigma_i^2.A_i^T \\ A_i.\Sigma_i^2 & A_i.\Sigma_i^2.A_i^T + \Phi_i^2 \end{pmatrix} \end{aligned}$$

Now let $Z = (X, Y)$ be the best Gaussian fit to the mixture of these two jointly Gaussian distributions. By the formulas above, we then have that

$$\begin{aligned} E[X] &= p_1\mu_1 \ + \ p_2\mu_2 \\ E[Y] &= p_1(A_1.\mu_1 + B_1) \ + \ p_2(A_2.\mu_2 + B_2) \end{aligned}$$

while the entries of the covariance matrix are

$$\begin{aligned} Cov[X, X] &= p_1(\Sigma_1^2) \ + \ p_2(\Sigma_2^2) \ + \\ &\qquad p_1p_2(\mu_1 - \mu_2).(\mu_1 - \mu_2)^T \\ Cov[Y, X] &= p_1(A_1.\Sigma_1^2) \ + \ p_2(A_2.\Sigma_2^2) \ + \\ &\qquad p_1p_2\,(A_1.\mu_1 + B_1 - A_2.\mu_2 - B_2).(\mu_1 - \mu_2)^T \\ Cov[Y, Y] &= p_1(A_1.\Sigma_1^2.A_1^T + \Phi_1^2) \ + \ p_2(A_2.\Sigma_2^2.A_2^T + \Phi_2^2) \ + \\ &\qquad p_1p_2\,(A_1.\mu_1 + B_1 - A_2.\mu_2 - B_2).(\mu_1 - \mu_2)^T \end{aligned}$$

Once these statistics are computed, a new slope and intercept can then be derived using the formulas $A = Cov[Y, X].Cov[X, X]^{-1}$ and $B = E[Y] - A.E[X]$.

3.4 The Ambiguity of Linear-Gaussian Mixing

As the formulas above indicate, the covariance $Cov[Y, X]$ of the mixture distribution depends on the covariance of the two component inputs,

$$Cov[X_1, X_1] = \Sigma_1^2 \qquad \text{and} \qquad Cov[X_2, X_2] = \Sigma_2^2.$$

Generally speaking, the most widely dispersed mixture component will also have the largest influence on $Cov[Y, X]$, reflecting the fact that a larger variety of input values leads to a more confident estimate of the covariance.

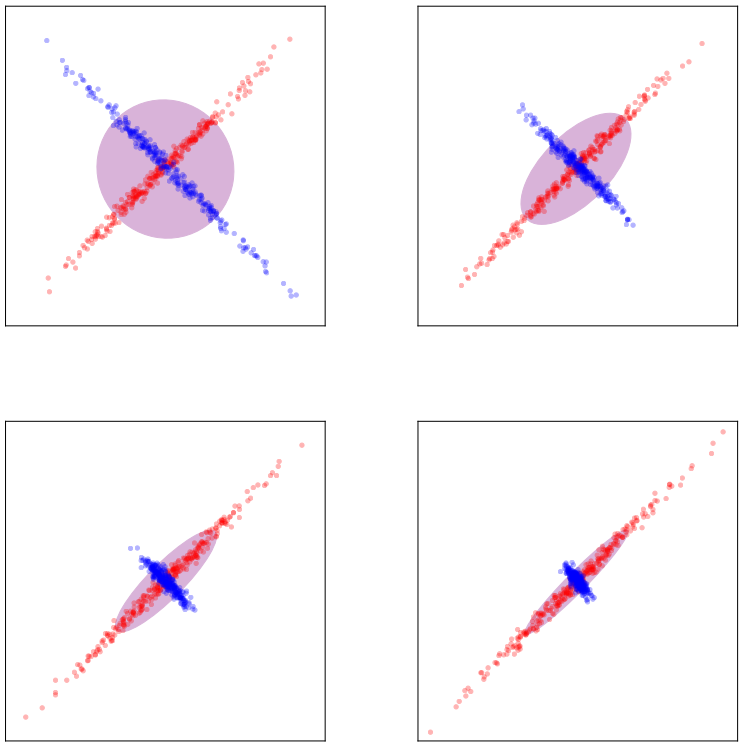

Figure 5: When two linear models are combined to form a third model, the slope of the resulting line depends heavily on the variance of the input fed into the two component models.

Consider for instance the two linear models

$$\begin{aligned} Y_1 &= +X_1 + \varepsilon_1 \\ Y_2 &= -X_2 + \varepsilon_2 \end{aligned}$$

with $X_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$ for $i = 1, 2$. Under these assumptions,

$$\begin{aligned} Cov[Y_1, X_1] &= +\sigma_1^2 \\ Cov[Y_2, X_2] &= -\sigma_2^2 \end{aligned}$$

The mixture variable (X, Y) with proportions $p_1 : p_2$ then satisfies

$$\begin{aligned} Cov[X, X] &= p_1\sigma_1^2 + p_2\sigma_2^2 \\ Cov[Y, X] &= p_1\sigma_1^2 - p_2\sigma_2^2 \end{aligned}$$

The slope estimate derived from this mixture is therefore

$$A = \frac{p_1\sigma_1^2 - p_2\sigma_2^2}{p_1\sigma_1^2 + p_2\sigma_2^2} \rightarrow \begin{cases} +1 & \text{for } (\sigma_1^2/\sigma_2^2) \rightarrow \infty \\ -1 & \text{for } (\sigma_1^2/\sigma_2^2) \rightarrow 0 \end{cases}$$

Figure 5 illustrates this idea for $p_1 = p_2 = 1/2$ and different values of the variance ratio σ_1^2/σ_2^2.

Hence, different assumptions about the input distribution lead to different best-fit lines, even when the input distribution distributions are fed into the same pair of linear models. This makes the choice of input distribution crucial in the cases where a jointly Gaussian model is not estimated directly, but derived by feeding a hypothetical input distribution into a linear function.