We are going to get into the mathematical principles behind completing the square, calculating conditional Gaussians, and Bayesian posterior calculation for linear Gaussian likelihoods.

There are two technical points, 1) a probability density function with linear and quadratic terms collecting in the exponent yields a (unnormalized) Gaussian distribution, and 2) once we collect the linear and quadratic terms of a certain variable we can analytically integrate out those terms.

1 The natural parameter and information form of a Gaussian

We derive the natural parameter form for a Gaussian and relate it to the completing the square form.

Going back to my alternative derivation of a Bernoulli, we can define an exponential family by defining the sufficient statistics and the intrinsic measure, $h(x)$.

It turns out we can treat the intrinsic measure as $h(x) = 1$, and the sufficient statistic vector as $u_1(x) = xx^\intercal$ and $u_2(x) = x$.

We can write the unnormalized probability distribution as $\tilde{p}(x|\eta_1, \eta_2) = h(x)e^{\langle xx^\intercal, \eta_1 \rangle + x^\intercal \eta_2}$. This means that when we see an unnormalized probability distribution written in terms of the quadratic term and linear term, it is close to the natural parameter form of a gaussian distribution.

We perform a more detailed derivation of the natural parameters in terms of the mean, μ, and the covariance, Σ.

We know that probability density functions integrate out to 1, which is equivalent to defining the normalization/partition function for the unnormalized distribution.

$$Z(\eta) := \int h(x)e^{u(x)^\intercal \eta} dx \tag{1}$$

For Gaussian distributions, we can rewrite the probability density in exponential family form

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{-\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)} \tag{2}$$

Choosing $u_1(x) = xx^\intercal$ and $u_2(x) = x$, we can collect the terms inside the exponential into quadratic, linear, and constant terms

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{(2\pi)^{-\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}x^\intercal \Sigma^{-1} x + \frac{1}{2}x^\intercal \Sigma^{-1}\mu + \frac{1}{2}\mu^\intercal \Sigma^{-1} x - \frac{1}{2}\mu^\intercal \Sigma^{-1}\mu} \\ &= \frac{1}{(2\pi)^{-\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\mu^\intercal \Sigma^{-1}\mu} e^{\left\langle xx^\intercal, -\frac{\Sigma^{-1}}{2} \right\rangle + \left\langle x, \Sigma^{-1}\mu \right\rangle} \end{aligned} \tag{3}$$

where in the second line we have used fact that Σ and Σ^{-1} are symmetric.

Then we see that our natural parameters are $\eta_1 = -\frac{\Sigma^{-1}}{2}$ and $\eta_2 = \Sigma^{-1}\mu$, and our normalization function is

$$\begin{aligned} Z(\mu, \Sigma) &= (2\pi)^{\frac{D}{2}} \det \Sigma^{\frac{1}{2}} e^{\frac{\mu^\intercal \Sigma^{-1} \mu}{2}} \\ Z(\eta_1, \eta_2) &= (2\pi)^{\frac{D}{2}} \det(-2\eta_1)^{-\frac{1}{2}} e^{\frac{\eta_2^\intercal (-2\eta_1)^{-1} \eta_2}{2}} \end{aligned} \tag{4}$$

This is a little unwieldy, so instead we can slightly modify the natural parameter form of a Gaussian to the information form (see Koller, Chapter 7, Gaussian Network Models), which is written in terms of the precision matrix, $\Lambda = \Sigma^{-1}$, and the mean.

$$p(\tilde{x}) = e^{\left\langle xx^\intercal, \frac{\Lambda}{2} \right\rangle + \langle x, \Lambda\mu \rangle} \tag{5}$$

Then we can write the normalization constant as

$$Z(\mu, \Lambda) = (2\pi)^{\frac{D}{2}} \det \Lambda^{-\frac{1}{2}} e^{\left\langle \mu\mu^\intercal, \frac{\Lambda}{2} \right\rangle} \tag{6}$$

and this lines up very well with our completing the square math.

1.1 Matrix dot products

We have rewritten our quadratic forms as a dot product between the matrix and an outer product of the two vectors, and here we justify this. The benefit of this matrix dot products form is that dot products are bilinear, allowing us to collect quadratic terms together:

$$\langle M_1, xx^\intercal \rangle + \langle M_2, xx^\intercal \rangle = \langle M_1 + M_2, xx^\intercal \rangle \tag{7}$$

To start, we observe what happens if we define a dot product between two matrices by vectorizing each matrix and then taking the dot product

$$\langle A, B \rangle = \sum_{i,j} A_{ij} B_{ij} \tag{8}$$

We can do something similar using matrix multiplication, though in order to multiply the j_1th column of A with j_2th column of B, we need to take the transpose of the j_1th column of A

$$\begin{aligned} \langle A_{\cdot j_1} B_{\cdot j_2} \rangle &= A_{\cdot j_1}^\intercal B_{\cdot j_2} \\ &= \sum_i A_{i, j_1} B_{i, j_2} \end{aligned} \tag{9}$$

We can break matrix dot product into two nested summations

$$\begin{aligned} \langle A, B \rangle &= \sum_j \sum_i A_{ij} B_{ij} \\ &= \sum_j A_{\cdot j}^\intercal B_{\cdot j} \end{aligned} \tag{10}$$

If we were to do a full matrix multiplication $A^\intercal B$, we note that the ij-th entry of the result is $A_{\cdot i}^\intercal B_{\cdot j}$. But we only are interested in the terms where $i = j$, that is the diagonal terms, and we want to sum them up.

The sum of the diagonal terms is defined exactly as the trace, so

$$\langle A, B \rangle = \operatorname{tr}(A^\intercal B) \tag{11}$$

Next we show that the quadratic form $x^\intercal M x$ can be written as $\langle M, xx^\intercal \rangle$, the matrix dot product of M and the outer product of x and itself.

Intuitively, we can view the $n \times x$ matrix M as something like a function defined over the product of two discrete spaces of size n. If we wanted to evaluate the expectation with respect to some probability distribution P over this joint space, we would simply multiply each term individually and add them together

$$\langle P, M \rangle = \sum_{i,j} P_{ij} M_{ij} \tag{12}$$

In the particular case where the marginal probabilities over each axis/discrete space are independent of each other, the joint probability can be written as the outer product of the marginal distributions: $P = P_I P_J^\intercal$.

Then we can marginalize out each axis separately

$$\begin{aligned} \langle P, M \rangle &= \sum_i \left(\sum_j M_{ij} P_j \right) P_i \\ &= P_i^\intercal M P_j \end{aligned} \tag{13}$$

Another way to see this is to note that the quadratic form $1_i^\intercal M 1_j$ is equal to M_{ij}, where 1_i and 1_j are one-hot vectors with 1's at the ith and jth entry, respectively.

Then we can view $x^\intercal M x$ as

$$\begin{aligned} x^\intercal M x &= \left(\sum_i x_i 1_i \right)^\intercal M \left(\sum_j x_j 1_j \right) \\ &= \sum_{ij} x_i x_j M_{ij} \\ &= \langle M, xx^\intercal \rangle \end{aligned} \tag{14}$$

2 Integration using the normalization function

In particular, we can integrate out our unnormalized $\tilde{p}(x|\eta_1, \eta_2) = e^{\langle xx^\intercal, \eta_1 \rangle + \langle x, \eta_2 \rangle}$ by setting it to the partition function

$$\begin{aligned} Z(\mu, \Lambda) &= \int e^{\langle xx^\intercal, \eta_1 \rangle + \langle x, \eta_2 \rangle} dx \\ &= (2\pi)^{\frac{D}{2}} \det \Lambda^{-\frac{1}{2}} e^{\frac{\mu^\intercal \Lambda \mu}{2}} \end{aligned} \tag{15}$$

3 Bayesian linear regression

With these additional theoretical insights, we are ready to deal with calculating the posterior of a linear Gaussian likelihood.

The steps are to 1) compute the joint probability of the weights and the targets, 2) collect the linear and quadratic terms in the exponent, 3) marginalize out the weights, which turns out to be the same as marginalizing out the posterior distribution, and 3) collecting the remaining linear and quadratic terms in the exponent into a gaussian distribution for the target.

$$\begin{aligned} &p(t|x,w,b) = \mathcal{N}\left(wx+b, \beta^{-1}\right) \\ &p(w) = \mathcal{N}\left(0, \alpha^{-1} I_{D_w}\right) \\ &p(w,t_1,\ldots,t_N|x_1,\ldots,x_N) = p(w)\prod_{n=1}^{N} p(t_n|x_n,w,b) \end{aligned} \tag{16}$$

The idea now is that we want to marginalize out w, by collecting the linear and quadratic terms of w in the exponent. Unfortunately, the exponential terms involving w also involve x and t, so we are not marginalizing out $p(w)$, we are marginalizing out the posterior distribution $p(w|x_1,\ldots,x_N,t_1,\ldots,t_N)$

$$p(w,t_1,\ldots,t_N|x_1,\ldots,x_N) = p(w|t_1,\ldots,t_N,x_1,\ldots,x_N)p(t_1,\ldots,t_N|x_1,\ldots,t_N) \tag{17}$$

Our integral over w is integrating the information form of a Gaussian distribution over w.

Then we can use our normalization function to evaluate the integral, and we argue that the leftover marginal distribution is once again in the information form of a Gaussian.

$$\begin{aligned} &\int p(w,x_1,\ldots,x_N,t_1,\ldots,t_N)dw = \\ &\int Z(\alpha)e^{-\frac{1}{2}w^\intercal \alpha I_{D_w} w}\prod_{n=1}^{N}(2\pi)^{-\frac{1}{2}}\beta^{\frac{1}{2}}e^{-\frac{1}{2}(t_n-(w\cdot x_n+b))^\intercal\beta(t_n-(w\cdot x_n+b))}dw = \\ &Z(\alpha)(2\pi)^{\frac{N}{2}}\beta^{\frac{N}{2}}\int e^{-\frac{1}{2}w^\intercal\alpha I_{D_w}w-\frac{1}{2}\sum_{n=1}^{N}t_n^\intercal\beta t_n+\sum_{n=1}^{N}t_n^\intercal\beta(w\cdot x_n+b)-\frac{1}{2}\sum_{n=1}^{N}(w\cdot x_n+b)^\intercal\beta(w\cdot x_n+b)}dw \end{aligned} \tag{18}$$

We're going to drop out the bias term b to make the math simpler - this would be equivalent to centering the data.

We collect the quadratic and linear w terms to put it in information form.

$$\begin{aligned} &Z(\alpha)(2\pi)^{\frac{N}{2}}\beta^{\frac{N}{2}}\int e^{-\frac{1}{2}\langle ww^\intercal,\alpha I_{D_w}\rangle-\frac{1}{2}\sum_{n=1}^{N}\langle ww^\intercal,\beta x_n x_n^\intercal\rangle+\sum_{n=1}^{N}\langle w,\beta t_n x_n\rangle-\frac{1}{2}\sum_{n=1}^{N}t_n^\intercal\beta t_n} \\ &= Z(\alpha)(2\pi)^{\frac{N}{2}}\beta^{\frac{N}{2}}e^{-\frac{1}{2}\sum_{n=1}^{N}t_n^\intercal\beta t_n}\int e^{-\frac{1}{2}\left\langle ww^\intercal,\alpha I_{D_w}+\beta\sum_{n=1}^{N}x_n x_n^\intercal\right\rangle+\left\langle w,\beta\sum_{n=1}^{N}t_n x_n\right\rangle}dw \end{aligned} \tag{19}$$

Thus the term inside the integral is equivalent to the normalization function of the unnormalized version of

$$\mathcal{N}\left(w; \mu_{w|\mathcal{D}} := \beta\Sigma_{w|\mathcal{D}} \sum_{n=1}^{N} t_n x_n, \Sigma_{w|\mathcal{D}} := \left(\alpha I_{D_w} + \beta \sum_{n=1}^{N} x_n x_n^\intercal\right)^{-1}\right) \tag{20}$$

This distribution is the posterior distribution, $p(w|\mathcal{D})$ conditioned on the dataset $\mathcal{D}$. Thus what we are doing by marginalizing out w from the joint is calculating the marginal distribution $p(\mathcal{D})$, using the posterior distribution.

The remaining terms are

$$Z(\alpha)\left(2\pi\right)^{\frac{N}{2}} \beta^{\frac{N}{2}} e^{-\frac{1}{2}\sum_{n=1}^{N} t_n^\intercal \beta t_n} \left(2\pi\right)^{-\frac{D}{2}} \det \Sigma_{w|\mathcal{D}}^{\frac{1}{2}} e^{\frac{1}{2}\mu_{w|\mathcal{D}}^\intercal \Sigma_{w|\mathcal{D}}^{-1} \mu_{w|\mathcal{D}}} \tag{21}$$

We note that in the exponent, everything is a quadratic term in terms of two t_n variables, and we are going to treat this is an unnormalized probability distribution over the vector t. In fact if we assume it marginalizes properly, we do not have to pay attention to the outside factors and just focus on the terms inside the exponential

$$\begin{aligned} -\frac{\beta}{2}\sum_{n=1}^{N} t_n^\intercal t_n + \left(\beta \sum_{n=1}^{N} t_n x_n\right)^\intercal \Sigma_{w|\mathcal{D}} \left(\beta \sum_{n=1}^{N} t_n x_n\right) = \\ -\frac{\beta}{2}\left\langle \vec{t}\vec{t}^\intercal, \beta I_N \right\rangle + \beta^2 \vec{t}^\intercal X \Sigma_{w|\mathcal{D}} X^\intercal \vec{t} \\ = -\frac{1}{2}\left\langle \vec{t}\vec{t}^\intercal, \beta I_N - \beta^2 X \Sigma_{w|\mathcal{D}} X^\intercal \right\rangle \end{aligned} \tag{22}$$

And finally we can use Woodbury's identity to recover ¡++¿

Therefore, (for centered data), the marginal target distribution has covariance ¡++¿ and is from a normal distribution $\mathcal{N}\left(0, < ++ >\right)$.