Notes for 751-752
Section 12

Martin Lindquist*

October 4, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

12 The normal distribution

12.1 The univariate normal distribution

A random variable Z follows a standard normal distribution if its density is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2).$$

We write the associated distribution function as F. A standard normal variate has mean 0 and variance 1. All odd numbered moments are 0. The non-standard normal variate, say X, having mean μ and standard deviation σ can be obtained as $X = \mu + \sigma Z$. Conversely, $(X - \mu)/\sigma$ is standard normal if X is any non-standard normal. The non-standard normal density is:

$$f\left(\frac{x-\mu}{\sigma}\right)/\sigma$$

with distribution function $F\left(\frac{x-\mu}{\sigma}\right)$.

12.2 The multivariate normal distribution

Suppose $Z_1, Z_2, \ldots Z_n$ are independent identically distributed (i.i.d.) standard normal random variables. The joint density of $\mathbf{z} = (Z_1, Z_2, \ldots Z_n)'$ is then given by

$$\begin{aligned} f_{\mathbf{z}}(\mathbf{z}) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2) \\ &= (2\pi)^{-p/2} \exp(-\mathbf{z}'\mathbf{z}/2) \end{aligned}$$

This is the multivariate standard normal distribution for a random vector $\mathbf{z}$ with mean $\mathbf{0}$ and variance $\mathbf{I}$. We write this as $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$.

Non standard normal variates, say $\mathbf{x}$, can be obtained as $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}$ where $E[\mathbf{x}] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$, which is assumed to be positive definite. Conversely, one can go backwards with $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. The non-standard multivariate normal distribution is given by

$$(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Commit this density to memory. In this setting, we say that $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The normal distribution is nice to work with in that all full row rank linear transformations of the normal are also normal. That is, if $\mathbf{a} + \mathbf{A}\mathbf{x}$ is normal if $\mathbf{A}$ is full row rank. Also, all conditional and submarginal distributions of the multivariate normal are also normal. (We'll discuss the conditional distribution more later.)

12.3 Moment generating functions

Given a vector $\boldsymbol{\mu}$ and a positive semidefinite matrix $\boldsymbol{\Sigma}$, $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if the moment generating function (m.g.f.) of $\mathbf{y}$ is

$$M_{\mathbf{y}}(\mathbf{t}) \equiv E[e^{\mathbf{t}'\mathbf{y}}] = \exp\{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\}.$$

Here are two important properties of moment generating functions:

- If two random vectors have the same moment generating function, they have the same density.
- Two random vectors are independent if and only if their joint moment generating function factors into the product of their two separate moment generating functions.

12.4 Properties of the multivariate normal distribution

Let $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \Sigma)$, and let $\mathbf{a}$ be a $p \times 1$ vector, $\mathbf{b}$ a $k \times 1$ vector, and $\mathbf{C}$ a $k \times p$ matrix with rank$= k \leq p$, then,

- $x = \mathbf{a}'\mathbf{y} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$.
- $\mathbf{x} = \mathbf{C}\mathbf{y} + \mathbf{b} \sim N_p(\mathbf{C}\boldsymbol{\mu} + \mathbf{b}, \mathbf{C}\Sigma\mathbf{C}')$.

Example: 12.1 Let $\mathbf{z} = (Z_1, Z_2)' \sim N_2(\mathbf{0}, \mathbf{I})$, and let $\mathbf{A}$ be the linear transformation matrix

$$\mathbf{A} = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Let $\mathbf{y} = (Y_1, Y_2)'$ be the linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{z} = \begin{pmatrix} (Z_1 - Z_2)/2 \\ (Z_2 - Z_1)/2 \end{pmatrix}.$$

Now $\mathbf{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$.

Theorem: 12.2 Let $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, and let $\mathbf{T}$ be an orthogonal constant matrix. Then $\mathbf{T}\mathbf{y} \sim N_n(\mathbf{T}\boldsymbol{\mu}, \sigma^2\mathbf{I})$.

Theorem 12.2 says that mutually independent normal random variables with common variance remain mutually independent with common variance under orthogonal transformations. Orthogonal matrices correspond to rotations and reflections about the origin, i.e., they preserve the vector length:

$$||\mathbf{Ty}||^2 = (\mathbf{Ty})'(\mathbf{Ty}) = \mathbf{y}'\mathbf{T}'\mathbf{Ty} = \mathbf{y}'\mathbf{y} = ||\mathbf{y}||^2.$$

Theorem: 12.3 If $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any $p \times 1$ subvector of $\mathbf{y}$ has a p-variate normal distribution.

It follows that if $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $Y_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots n$. Thus, joint normality implies marginal normality. The converse is not necessarily true.

12.5 Singular normal

What happens if the $\mathbf{A}$ in the previous section is not of full row rank? Then Var($\mathbf{X}$) $=$ $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ is not full rank. There are redundant elements of the vector $\mathbf{X}$ in that if you know some of them, you know the remainder.

An example is our residuals. The matrix $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})$ is not of full rank (it's rank is $n - p$). For example, if we include an intercept, the residuals must sum to 0. Know any $n-1$ of them and you know the n^{th}. A contingency for this is to define the singular normal distribution. A singular normal random variable is any random variable that can be written as $\mathbf{Az} + \mathbf{b}$ for a matrix $\mathbf{A}$ and vector $\mathbf{b}$ and standard normal vector $\mathbf{z}$. As an example, consider the case where $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Then the residuals, defined as

$$\{\mathbf{I} - \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}\}\mathbf{y} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}\}(\mathbf{X}\boldsymbol{\beta} + \frac{1}{\sigma}\mathbf{z})$$

are a linear transformation of iid normals. Thus the residuals are singular normal.

The singular normal is such that all linear combinations and all submarginal and conditional distributions are also singular normal. The singular normal doesn't necessarily have a density function, because of the possibility of redundant entries. For example, the vector $(Z\ Z)$, where Z is a standard normal, doesn't have a joint density since the covariance matrix is $\mathbf{1}_{2\times 2}$, which isn't invertible.

The normal is the special case of the singular normal where the covariance matrix is full rank.

12.6 Independence

Let $\sim \mathbf{N_n}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be partitioned as follows:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where $\mathbf{y}_1$ is $p \times 1$ and $\mathbf{y}_2$ is $q \times 1$, $(p+q=n)$. Then, the mean and covariance matrix are correspondingly partitioned as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{Y}_1) & \text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) \\ \text{cov}(\mathbf{Y}_2, \mathbf{Y}_1) & \text{var}(\mathbf{Y}_2) \end{pmatrix}.$$

The marginal distributions are $\mathbf{Y}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{Y}_2 \sim N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

If

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

is $N_{p+q}(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent if $\Sigma_{12} = \mathbf{0}$. However, if $\mathbf{Y}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{Y}_2 \sim N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \mathbf{0}$, this does not necessarily mean that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent. We also need $\mathbf{Y}$ to be jointly normal. It follows that if $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then any two individual variables y_i and y_j are independent if $\sigma_{ij} = 0$.

12.7 Conditional distributions

The conditional distribution of a normal is often of interest. Let $\mathbf{X} = [\mathbf{X}_1' \ \mathbf{X}_2']'$ be comprised of an $n_1 \times 1$ and $n_2 \times 1$ matrix where $n_1 + n_2 = n$. Assume that $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1' \ \boldsymbol{\mu}_2']'$ and

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Consider now the conditional distribution of $\mathbf{X}_1 \mid \mathbf{X}_2$. A clever way to derive this (shown to me by a student in my class) is as follows let $\mathbf{Z} = \mathbf{X}_1 + \mathbf{A}\mathbf{X}_2$ where $\mathbf{A} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. Then note that the covariance between $\mathbf{X}_2$ and $\mathbf{Z}$ is zero (HW).

Thus the distribution of $\mathbf{Z} \mid \mathbf{X}_2$ is equal to the distribution of $\mathbf{Z}$ and that it is normally distributed being a linear transformation of normal variates. Thus we know both

$$E[\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2] + \mathbf{A}E[\mathbf{X}_2 \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2] + \mathbf{A}\mathbf{x}_2$$

and

$$E[\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2] = E[\mathbf{Z}] = \boldsymbol{\mu}_1 + \mathbf{A}\boldsymbol{\mu}_2.$$

Setting these equal we get that

$$E[\mathbf{X}_1 \mid \mathbf{X}_2] = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

As a homework, using the same technique to derive the conditional variance

$$\text{Var}(\mathbf{Z} \mid \mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'.$$

12.8 An important example

Consider the vector $(\mathbf{Y}\ \mathbf{X}')'$ where $\mathbf{Y}$ is 1×1 and $\mathbf{X}$ is $p \times 1$. Assume that the vector is normal with $E[\mathbf{Y}] = \mu_y$, $E[\mathbf{X}] = \boldsymbol{\mu}_x$ and the variances are σ_y^2 (1×1) and $\boldsymbol{\Sigma}_x$ $(p \times p)$ and covariance $\boldsymbol{\rho}_{xy}$ $(p \times 1)$.

Consider now predicting $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Clearly a good estimate for this would be $E[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$. Our results suggest that $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$ is normal with mean:

$$\mu_y + \boldsymbol{\rho}_{xy}' \boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) = \mu_y - \boldsymbol{\mu}_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\rho}_{xy} + \mathbf{x}' \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\rho}_{xy} = \beta_0 + \mathbf{x}' \boldsymbol{\beta}$$

where $\beta_0 = \mu_y - \boldsymbol{\mu}_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\rho}_{xy}$ and $\beta = \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\rho}_{xy}$. That is, the conditional mean in this case mirrors the linear model. The slope is defined exactly as the inverse of the variance/covariance matrix of the predictors times the cross correlations between the predictors and the response. We discussed the empirical version of this in Section **??** where we saw that the empirical coefficients are the inverse of the empirical variance of the predictors times the empirical correlations between the predictors and response. A similar mirroring occurs for the intercept as well.

This correspondence simply says that empirical linear model estimates mirror the population parameters if both the predictors and response are jointly normal. It also yields a motivation for the linear model in some cases where the joint normality of the predictor and response is conceptually reasonable. Though we note that often such joint normality is not reasonable, such as when the predictors are binary, even though the linear model remains well justified.

12.9 A second important example

Consider our partitioned variance matrix.

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma} 12' & \boldsymbol{\Sigma}_{22} \end{array} \right].$$

The upper diagonal element of $\boldsymbol{\Sigma}^{-1}$ is given by $\mathbf{K}_{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12}')^{-1}$, which is the inverse of $\text{Var}(\mathbf{X}_1 \mid \mathbf{X}_2)$. Suppose that $\mathbf{X}_1 = (X_{11}\ X_{12})'$. Then this result suggests that X_{11} is independent of X_{12} given $\mathbf{X}_2$ if the $(1,2)$ off diagonal element of $\boldsymbol{\Sigma}^{-1}$ is zero. To see this, recall that independence and absence of correlation are equivalent in the multivariate normal. Hence, if X_{11} is independent of X_{12} given $\mathbf{X}_2$ then $\mathbf{K}_{11}^{-1}$ is diagonal, and so must $\mathbf{K}_{11}$. There's nothing in particular about the first two positions, so we arrive at the following remarkable fact: whether or not the off diagonal elements of $\boldsymbol{\Sigma}^{-1}$ are zero determines the conditional independence of those random variables given the remainder. This forms the basis of so-called Gaussian graphical models. The graph defined by ascertaining which elements of $\boldsymbol{\Sigma}^{-1}$ are zero is called a conditional independence graph.

12.10 Normal likelihood

Let $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ then note that minus twice the log-likelihood is:

$$n\log(\sigma^2) + ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2/2\sigma^2$$

Holding σ^2 fixed we see that minmizing minus twice the log likelihood (thus maximizing the likelihood) yields the least squares solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Since this doesn't depend on σ it is the MLE. Taking derivatives and setting equal to zero we see that

$$\hat{\sigma}^2 = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2/n$$

(i.e. the average of the squared residuals). We'll find that there's a potentially preferable unbiased estimate given by

$$s^2 = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2/(n-p).$$

This model can be written in a likelihood equivalent fashion of

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$. However, one must be careful with specifying linear models this way. For example, if one wants to simulate $Y = X + Z$ where X and Z are generated independently, one can not equivalently simulate X by generating Y and Z independently and taking $Z - Y$. (Note Y and Z are correlated in the original simulation specification.) Writing out the distributions explicitly removes all doubt. Thus the linear notation, especially when there are random effects, is sort of lazy and imprecise (though everyone, your author included, uses it).

Let's consider another case, suppose that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are iid p vectors $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, disregarding constants, minus twice the log likelihood is

$$n\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}).$$

Assume that $\boldsymbol{\Sigma}$ is known, then using our derivative rules from earlier, we can minimize this to obtain the MLE for $\boldsymbol{\mu}$

$$\hat{\mu} = \bar{\mathbf{y}}$$

and the following for $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$$

Consider yet another case $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with known $\boldsymbol{\Sigma}$. Minus twice the log-likelihood is:

$$\log|\boldsymbol{\Sigma}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Using our matrix rules we find that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}.$$

This is the so-called generalized least squares estimate.

12.11 Bayes calculations

We assume a slight familiarity of Bayesian calculations and inference for this section. In a Bayesian analysis, one multiplies the likelihood times a prior distribution on the parameters to obtain a posterior. The posterior distribution is then used for inference. Let's go through a simple example. Suppose that $\mathbf{y} \mid \mu \sim N(\boldsymbol{\mu}\mathbf{1}_n, \sigma^2 I)$ and $\mu \mid N(\mu_0, \tau^2)$ where $\mathbf{y}$ is $n \times 1$ and μ is a scalar. The normal distribution placed on μ is called the "prior" and μ_0 and τ^2 are assumed to be known. For this example, let's assume that σ^2 is also known. The goal is to calculate $\mu \mid \mathbf{y}$, the posterior distribution. This is done by multiplying prior times likelihood. Since, operarting generically,

$$f(\text{Param}|\text{Data}) = \frac{f(\text{Param}, \text{Data})}{f(\text{Data})} \propto f(\text{Data}|\text{Param})f(\text{Param}) = \text{Likelihood} \times \text{Prior}.$$

Here, the proportional symbol, $\propto$, is with respect to the parameter.

Consider our problem, retaining only terms involving μ we have that minus twice the natural log of the distribution of $\mu \mid \mathbf{y}$ is given by

$$\begin{aligned}
& -2\log(f(\mathbf{y} \mid \mu)) - 2\log(f(\mu)) \\
= \; & ||\mathbf{y} - \mu\mathbf{1}_n||^2/\sigma^2 + (\mu - \mu_0)^2/\tau^2 \\
= \; & -2\mu n\bar{y}/\sigma^2 + \mu^2 n/\sigma^2 + \mu^2/\tau^2 - 2\mu\mu_0/\tau^2 \\
= \; & -2\mu\left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right) + \mu^2\left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right)
\end{aligned}$$

This is recognized as minus twice the log density of a normal distribution for μ with variance of

$$\text{Var}(\mu \mid \mathbf{y}) = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right)^{-1} = \frac{\tau^2\sigma^2/n}{\sigma^2/n + \tau^2}$$

and mean of

$$E[\mu \mid \mathbf{y}] = \left(\frac{1}{\sigma^2/n} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{\bar{y}}{\sigma^2/n} + \frac{\mu_0}{\tau^2}\right) = p\bar{y} + (1-p)\mu_0$$

where

$$p = \frac{\tau^2}{\tau^2 + \sigma^2/n}.$$

Thus $E[\mu \mid \mathbf{y}]$ is a mixture of the empirical mean and the prior mean. How much the means are weighted depends on the ratio of the variance of the mean (σ^2/n) and the prior variance (τ^2). As we collect more data ($n \rightarrow \infty$), or if the data is not noisy ($\sigma \rightarrow 0$) or we have a lot of prior uncertainty ($\tau \rightarrow \infty$) the empirical mean dominates. In contrast as we become more certain a priori ($\tau \rightarrow 0$) the prior mean dominates.