Some Notes on Gaussian Processes & Regression

RLW, Draft 2.1

March 20, 2014

1 Positive Definite Matrices

Let $Y = (Y_1, \cdots, Y_p)'$ be a p-dimensional random vector, thought of as a random $(p \times 1)$ matrix. If they exist, the mean vector and covariance matrix for Y are

$$\mu := \mathsf{E}Y \qquad \Sigma := \mathsf{E}(Y-\mu)(Y-\mu)'$$

(they will both exist whenever $\mathsf{E}(Y_j)^2 < \infty$ for each j). The entries μ_i of the $p \times 1$ column matrix μ are the means $\mu_i = \mathsf{E}X_i$ of the variables, while the diagonal entries Σ_{ii} of the $p \times p$ matrix Σ are the variances of the $\{Y_i\}$. These diagonal entries must be nonnegative, but the off-diagonal entries

$$\Sigma_{ij} = \mathsf{E}(Y_i - \mu_i)(Y_j - \mu_j)$$

may be positive, negative, or zero as Y_i and Y_j are positively-, negatively-, or un-correlated. These entries aren't entirely arbitrary, though; clearly Σ must be symmetric, but even that isn't enough.

For any p real numbers $\{z_j\}$ the square of the linear combination $z'(Y-\mu) = \sum z_j(Y_j - \mu_j)$ is positive— so necessarily its expectation is too:

$$0 \le \mathsf{E}\left|\sum_j z_j(Y_j - \mu_j)\right|^2 = \mathsf{E}[z'(Y-\mu)][z'(Y-\mu)]' = z'\Sigma z = \sum_{ij} z_i \Sigma_{ij} z_j,$$

a property called "positive semi-definiteness" which implies that Σ has nonnegative eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$ with linearly-independent corresponding eigenvectors $\{u_j\} \subset \mathbb{R}^p$. If the distribution is nonsingular (so the $\{Y_i\}$ are linearly independent) then $z'\Sigma z > 0$ for $0 \ne z \in \mathbb{R}^p$, each $\lambda_j > 0$ is *strictly* positive, and Σ is called "positive-definite". The cone of all positive-definite matrices, denoted $\mathcal{S}_+^p$, is closed under addition $\Sigma_1 + \Sigma_2$ and conjugation $Q'\Sigma Q$ by nonsingular $p \times p$ matrices Q, and in particular under multiplication by positive scalars. SO, if Σ_1 and Σ_2 are positive-definite and $a_1, a_2 > 0$ then $a_1\Sigma_1 + a_2\Sigma_2 \in \mathcal{S}_+^p$ is also a positive-definite covariance matrix.

Given any vector $\mu \in \mathbb{R}^p$, and any positive-definite matrix $\Sigma \in \mathcal{S}_+^p$, there exists a Gaussian random vector $Y \sim \mathsf{No}(\mu, \Sigma)$ with mean μ and covariance Σ, and joint pdf

$$f(y \mid \mu, \Sigma) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\}.$$

One way to construct it is to begin with a vector Z of p independent standard normal random variables $Z_j \sim \mathsf{No}(0,1)$, take the eigendecomposition $\Sigma = U\Lambda U'$ of Σ into the product of a unitary matrix U (whose columns $\{u_j\}$ are unit eigenvectors of Σ), a diagonal matrix Λ (whose entries are the corresponding eigenvalues $\{\lambda_j\}$), and the transpose U', then set

$$Y := \mu + U\Lambda^{\frac{1}{2}}Z$$

and note Y is a linear combination of normals and hence is normal, with mean $\mathsf{E}[Y] = \mu$ and (since $\mathsf{E}ZZ' = I$, the $p \times p$ identity matrix) covariance

$$\begin{aligned}\mathsf{E}[(Y-\mu)(Y-\mu)'] &= \mathsf{E}\left[U\Lambda^{\frac{1}{2}}Z \quad Z'\Lambda^{\frac{1}{2}\prime}U'\right] \\ &= U\Lambda^{\frac{1}{2}} \quad I \quad \Lambda^{\frac{1}{2}}U' \\ &= U\Lambda U' = \Sigma.\end{aligned}$$

Another is to begin with the Cholesky decomposition $\Sigma = LL'$ and set $Y := \mu + LZ$, with mean $\mathsf{E}Y = \mu$ and covariance $\mathsf{E}[LZZ'L'] = LL' = \Sigma$.

2 Positive Definite Functions

If we have infinitely-many random variables $\{Y_i\}$ we need to do something a bit different, since joint pdfs won't exist (there is no Lebesgue measure "dx" in $\mathbb{R}^\infty$). Typically each observation Y_i is associated with a (vector of) covariate(s) x_i, like "time" or "location," which will affect the mean and covariance. Let $\mathcal{X}$ denote the set where these covariates lie (often $\mathbb{R}^q$ or some simple subset of it, for some integer $q \in \mathbb{N}$). The $\{Y_i\}$ are called *jointly Gaussian*, and $\{Y\}$ a "Gaussian Process" or "GP", if for each **finite** set I of indices the vector $Y_I = \{Y_i : \ i \in I\}$ is multivariate normal; in that case the joint distribution of all $\{Y_i\}$ is completely determined by the mean *function* and covariance *function*

$$\mu(x_i) = \mathsf{E}Y_i \qquad C(x_i, x_j) = \mathsf{E}[Y_i - \mu(x_i)][Y_j - \mu(x_j)]' \qquad x_i, x_j \in \mathcal{X}.$$

Thus each univariate marginal distribution is normal, so each $Y_i \sim \mathsf{No}(\mu_i, \sigma_i^2)$ has a normal distribution with mean $\mu_i = \mu(x_i)$ and variance $\sigma_i^2 = C(x_i, x_i)$, and each vector pair $[Y_i, Y_j]'$ has a bivariate normal distribution with mean $[\mu(x_i), \mu(x_j)]'$ and covariance $\Sigma_{ij} = C(x_i, x_j)$.

The mean function $\mu(x)$ for a Gaussian Process is completely arbitrary, but the covariance must always be positive-definite— *i.e.*, for each $p \in \mathbb{N}$ and $\{x_j\} \subset \mathcal{X}^p$,

$$0 < \sum_{ij} z_i C(x_i, x_j) z_j$$

for each $0 \neq z \in \mathbb{R}^p$. Just as covariance *matrices* form a positive cone $\mathcal{S}_+^p$, so do covariance *functions*— so, if C_1 and C_2 are positive definite, and $a_1, a_2 > 0$ constant, then $a_1 C_1(x,y) + a_2 C_2(x,y)$ is a positive-definite function too and is the covariance function for some Gaussian process.

The covariance function C is called "stationary" if it is translation-independent, *i.e.*, if $C(x_1, x_2)$ depends only on the vector difference $(x_1 - x_2)$, and "isotropic" if it depends only on the (usually Euclidean) distance $|x_1 - x_2|$. Commonly used families of isotropic covariance functions include the Matérn family, the power-exponential, and spherical. Not all GP covariance functions *are* isotropic or stationary, though, and not all data are well represented by isotropic GPs. If data are heteroskedastic (with higher variance for some $x \in \mathcal{X}$ than for others), for example, or if differences $[Y_i - Y_j]$ have variances that depend not only on distances $|x_i - x_j|$ but also the directions $(x_i - x_j)$ or even more general dependence on (x_i, x_j), then to use Gaussian Processes one must either (1) use a GP with a non-isotropic or non-stationary covariance, or (2) find transformations of $\mathcal{X}$ or an x-dependent transformation of Y to achieve homoskedacity and near-isotropy.

2.1 Example 1: Regression

Let $\mathcal{X} = [-1, 1]$ be an interval, let $a_i \overset{\text{ind}}{\sim} \mathsf{No}(\mu_i, \sigma_i^2)$ be independent normal random variables for $i = 0, 1, 2$, and set

$$Y(x) = a_0 + a_1 x + a_2 x^2.$$

Then $\{Y(x)\}$ is a GP with mean

$$\begin{aligned} \mu(x) &= \mathsf{E}[a_0 + a_1 x + a_2 x^2] \\ &= \mu_0 + \mu_1 x + \mu_2 x^2 \end{aligned}$$

and covariance function

$$\begin{aligned} C(x, y) &= \mathsf{E}\Big\{ \big[(a_0 - \mu_0) + (a_1 - \mu_1)x + (a_2 - \mu_2)x^2\big] \times \\ &\qquad \big[(a_0 - \mu_0) + (a_1 - \mu_1)y + (a_2 - \mu_2)y^2\big] \Big\} \\ &= \sigma_0^2 + xy\sigma_1^2 + x^2y^2\sigma_2^2 \end{aligned} \tag{1}$$

for $-1 \le x, y \le 1$. Figure 1 shows a plot of ten realizations with each $\mu_j = 0$ and $\sigma_j = 1$. Note that $C(x, y)$ can be *negative* for some x, y if $\sigma_1^2 > \sigma_0^2 + \sigma_2^2$; for example, if $\sigma_0 = \sigma_2 = 1$ while $\sigma_1 = 2$, then $C(-1, +1) = -2$ so $Y(-1)$ and $Y(+1)$ have covariance -2.

Generalizing the covariance of Eqn (1), for any number d of regression functions $\{\psi_m\}$ we can take $a_m \overset{\text{ind}}{\sim} \mathsf{No}(\mu_m, \sigma_m^2)$ and construct a GP $Y(x) = \sum_{0 \le m < d} a_m \psi_m(x)$ with mean $\mu(x) = \sum \mu_m \psi_m(x)$ and covariance $C(x, y) = \sum \sigma_m^2 \psi_m(x) \psi_m(y)$. The illustration above took $d = 3$ and $\psi_m(x) = 1, x, x^2$, but any number d of basis functions $\{\psi_m\}$ can work.

2.2 Example 2: Power-Exponential

For some applications we expect Y to vary continuously with varying x, suggesting we should want $Y(x_i)$ and $Y(x_j)$ to be nearly equal when $|x_i - x_j|$ is small and nearly independent when $|x_i - x_j|$ is large. Here's a model to accomplish that.

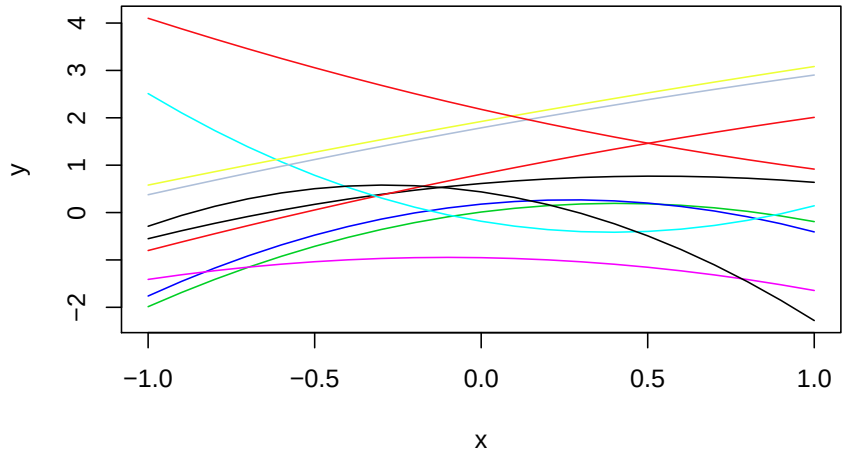

Figure 1: Sample paths for quadratic regression model

For a covariate space $\mathcal{X} \subset \mathbb{R}^q$ of q dimensions, fix $2q+1$ parameters:

Variance	$\sigma^2 > 0$	Marginal variance for each $Y(x)$
Shapes	$\vec{\alpha} \in [1,2]^q$	Determines smoothness in each coordinate
Scales	$\vec{\ell} \in \mathbb{R}_+^q$	Length scale (could be different in each dimension)

These determine the stationary (but not isotropic, unless the α's and ℓ's are constant) covariance function

$$C(x,y) = \sigma^2 \exp\left\{ -\sum_{k=1}^{q} \left| \frac{x_k - y_k}{\ell_k} \right|^{\alpha_k} \right\}, \tag{2}$$

a positive-definite function with maximum value σ^2 when $x = y \in \mathbb{R}^q$ that falls off exponentially in $(x_k/\ell_k)^{\alpha_k}$ in the kth direction. Sample-paths are always continuous but range from very rough with $\alpha_k = 1$ to very smooth with $\alpha_k = 2$. Figure 2 shows ten sample-paths for a GP with mean $\mu(x) = \sin(3x)$, $\sigma = 1/4$, smoothness $\alpha = 2$, and scale $\ell = 1/4$ (the mean is shown as a dashed black curve). This is the most commonly-used covariance in modeling computer model output, usually with $\alpha_k = 2$ (or slightly smaller to improve numerical stability), often with an added nugget term (again for stability).

Some readers (and reviewers), mesmerized by the similarity of Eqn (2) to the normal probability density function, confuse the role of "Gaussian" in a GP, thinking it describes the form of the covariance function $C(x,y)$ instead of the distribution of the $\{Y(x_i)\}$. The covariance function can be any positive-definite function— which does *not* require that $C(x,y) \geq 0$ for all x, y, as we have seen already with Eqn (1).

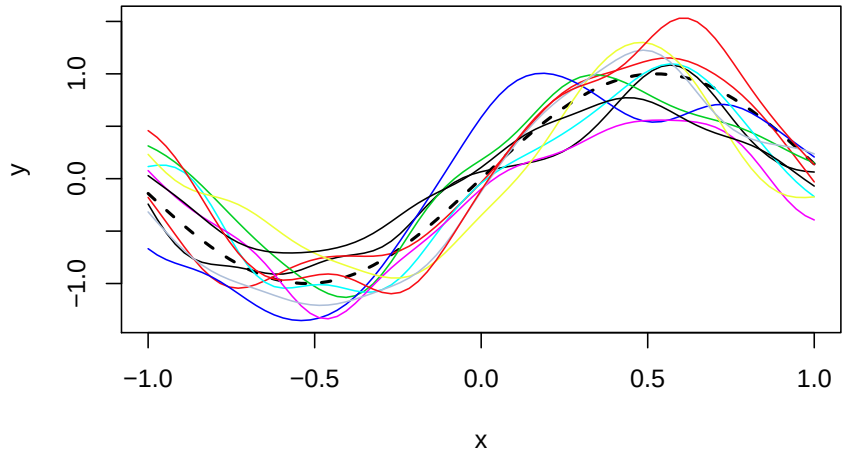

Figure 2: Sample paths for power-exponential model

2.3 Example 3: Hybrid of Regression + Power-Exponential

Sample paths for the model of Section 2.1 (shown in Figure 1) are all perfect parabolas, with no "noise", and with strong (positive or negative) correlation for $Y(x)$ at points at $x = +1$ and $x = -1$ for some values of $\{\sigma_i\}$. Sample paths for the model of 2.2 (shown in Figure 2) are smooth but feature negligible correlation for distant points. Some applications might feature both broad features (like the slope and curvature of the quadratic models of Figure 1) that persist across all observations, along with local wiggling (like that of Figure 2). These can all be accommodated in a GP with positive-definite covariance of the form

$$C(x_i, x_j) = \sum_{0 \le m < d} \sigma_m^2 \psi_m(x_i) \psi_m(x_j) + \sigma_d^2 \exp\left\{ - \sum_{k=1}^{q} \left| \frac{x_{ik} - x_{jk}}{\ell_k} \right|^{\alpha_k} \right\}.$$

If there is additional measurement-error or replication variation associated with individual *observations* and not just locations (in computer experiments this typically arises only when stochastic methods are used in model evaluation, but it always happens for observations of field data), a so-called "nugget" $\sigma_n^2 \delta_{ij}$ proportional to the Kronecker delta can be added to $C(x_i, x_j)$ above. Without this nugget the posterior distribution for all these curves will exactly duplicate the computer model output at the design points where the model was run; with a nugget, the curves will only come close to the computer outputs (about $\pm\sigma_n$, typically). By increasing the diagonal elements of the covariance matrix, adding such a nugget will improve its conditioning and improve numerical stability.

3 Connections with PCA

Computing unconditional samples from the models of Section 2.2 (as in Figure 2) or conditional ones (given observed values of a computer model actually evaluating Y_i at a few values of x_i) at some number p of points entails the inversion of a $p \times p$ matrix, requiring $O(p^3)$ operations and $O(p^3)$ storage locations, making this impractical for large p. One solution is to reduce the dimension to a smaller number $r \ll p$ by some means.

Consider now the case where $\mathcal{X} \subset \mathbb{R}^q$ is compact and $C(x, y)$ is continuous and hence bounded on $\mathcal{X}^2$. Then the Fredholm integral operator on $L_2(\mathcal{X}, dx)$ given by

$$\mathfrak{C}[f](x) = \int_{\mathcal{X}} C(x, y) f(y)\, dy$$

is positive and trace-class, since

$$\langle f, \mathfrak{C}[f] \rangle = \int_{\mathcal{X}^2} f(x)\, C(x, y) f(y)\, dx\, dy \geq 0$$

by the positive-definiteness of C. It follows that $\mathfrak{C}$ has orthonormal eigenfunctions $\{\phi_n\} \subset L_2(\mathcal{X}, dx)$ with summable nonnegative eigenvalues $\sigma_n^2 > 0$, and that we have the representations

$$Y(x) = \sum_{n<\infty} \sigma_n \phi_n(x) Z_n \qquad\qquad C(x, y) = \sum_{n<\infty} \sigma_n^2 \phi_n(x) \phi_n(y)$$

for iid $\{Z_n\} \overset{\text{iid}}{\sim} \mathsf{No}(0, 1)$ (this goes by the name of "Karhunen-Loève expansion"). For any $r \in \mathbb{N}$ we can consider the "reduced" process Y_r with covariance C_r given by

$$Y_r(x) = \sum_{n \leq r} \sigma_n \phi_n(x) Z_n \qquad\qquad C_r(x, y) = \sum_{n \leq r} \sigma_n^2 \phi_n(x) \phi_n(y).$$

The processes Y and Y_r will differ in squared L_2 norm by $\sum_{n>r} \sigma_n^2$, a fraction $\sum_{n>r} \sigma_n^2 / \sum_n \sigma_n^2$ of the entire variance (note $\sum_n \sigma_n^2 = \int_{\mathcal{X}} C(x, x)\, dx < \infty$).

Thus every GP on a compact $\mathcal{X} \subset \mathbb{R}^q$ is either a finite or an infinite version of the regression model of Section (2.1).

3.1 Practical Matters

In practice we don't know $C(x, y)$ exactly, hence can't find $\{\phi_n\}$ or $\{\sigma_n^2\}$, and we only observe $Y(x)$ at a few locations x_i. Still, those values determine a positive-definite covariance matrix

$$\Sigma_{ij} = C(x_i, x_j)$$

whose (largest few, if necessary) eigenvalues σ_n^2 and eigenvectors u_n could be discerned or approximated if we only knew Σ_{ij}. We don't, of course, but we can approximate it and

$\mu_i = \mu(x_i)$ by sample estimates

$$\hat{\mu}_i = \frac{1}{N}\sum_n Y_n(x_i) \qquad\qquad \hat{\Sigma}_{ij} = \frac{1}{N}\sum_n \big(Y_n(x_i) - \hat{\mu}_i\big)\big(Y_n(x_j) - \hat{\mu}_j\big)$$

based on N replicates, leading through eigendecomposition $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}'$ with $\Lambda = \mathrm{diag}(\{\hat{\sigma}_n^2\})$ to representations

$$Y_r(x_i) = \sum_{n \le r} \hat{U}_{in}\hat{\sigma}_n Z_n \qquad\qquad C_r(x_i, x_j) = \sum_{n \le r} \hat{U}_{in}\hat{\sigma}_n^2 \hat{U}_{jn}.$$

4 Posterior Distributions

In our applications Y_i represents the output of a computer model at input vector x_i; the input vector may include both components whose value we observe or even specify (called "environmental" and "control" variables in the literature: Sacks et al., 1989, §2.1), as well as uncertain features that we would like to discover ("model" variables). The computer model is too slow or expensive to evaluate it at an exhaustive collection of inputs, so we instead evaluate it at a set of carefully chosen "design" points $x_\mathcal{D} = \{x_i\}_{i\in\mathcal{D}} \subset \mathcal{X}$ (often chosen from a space-filling LHC design), often with the goal of predicting the values $Y_\mathcal{P} = \{Y_i\}_{i\in\mathcal{P}}$ at other sites $x_\mathcal{P} = \{x_i\}_{i\in\mathcal{P}}$ where the computer model has not been run.

If we use a GP for modeling the *prior* distribution, describing how uncertain Y_i is at each x_i before introducing as data the output of the computer model, then the conditional distribution of $Y_\mathcal{P}$ given the observed values $Y_\mathcal{D}$ at the design points (the *posterior* distribution) is also a GP, but now the mean $\mu(x)$ and covariance functions become

$$\mu_{\mathcal{P}|\mathcal{D}} = \mu_\mathcal{P} + \Sigma_{\mathcal{P}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\big[Y_\mathcal{D} - \mu_\mathcal{D}\big] \qquad \Sigma_{\mathcal{P}\mathcal{P}|\mathcal{D}} = \Sigma_{\mathcal{P}\mathcal{P}} - \Sigma_{\mathcal{P}\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}\mathcal{P}}$$

where, for example, $\Sigma_{\mathcal{D}\mathcal{P}}$ has entries $\mathsf{E}[(Y_i - \mu_i)(Y_j - \mu_j)]$ for $i \in \mathcal{D}$, $j \in \mathcal{P}$. Figure 3 shows 50 sample paths from this conditional distribution, given observed values at four design points $x_\mathcal{D} = \{-0.75, -0.5, +0.25, +0.5\}$. All paths pass through the design points exactly, and variability is small near those points and large away from them.

4.1 Parameter Space "Reduction"

The fifty posterior curves in Figure (3) are more tightly arrayed and fill a smaller area than the ten prior curves in Figure (2); in some sense the posterior is more concentrated than the prior. In this section we will quantify how the volume of plausible values for the parameters shrinks with the observation of data.

4.1.1 Conjugate Prior Regression

The standard Gaussian linear regression model is

$$Y \mid X \sim \mathsf{No}(X\beta, \Sigma)$$

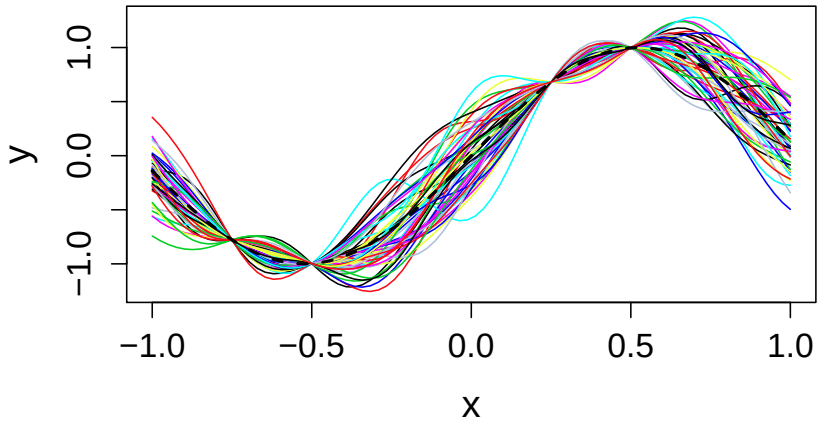

Figure 3: Posterior sample paths for power-exponential model, given $Y_{\mathcal{D}}$ at four design points $x_{\mathcal{D}}$.

for some $n \times p$ design (or covariate) matrix X, $n \times n$ error covariance matrix Σ (often, but not necessarily, diagonal or even of the form $\sigma^2 I$ for iid "errors"), and uncertain regression vector $\beta \in \mathbb{R}^p$. For full-rank X the maximum likelihood estimator $\hat{\beta}$ for β is a linear combination of the $\{Y_i\}$, hence once again normally distributed:

$$\begin{aligned} \hat{\beta} &= [X'\Sigma^{-1}X]^{-1}X'\Sigma^{-1}Y \\ &\sim \mathsf{No}\big(\beta, [X'\Sigma^{-1}X]^{-1}\big). \end{aligned}$$

In the Bayesian approach, β (like all uncertain quantities) is viewed as a random variable with some *prior* probability distribution. The most convenient choice is to take $\beta \sim \mathsf{No}(b, V)$ to be normal too, with some "prior mean" vector $b \in \mathbb{R}^q$ and "prior covariance" matrix $V \in \mathcal{S}_+^p$. The Bayes estimate is mean of the *posterior* (*i.e.*, conditional on Y) distribution of β; with a Gaussian prior distribution, this posterior distribution is once again Gaussian (that's what makes this choice so convenient) with posterior mean:

$$b_Y = \mathsf{E}[\beta \mid Y] \qquad\qquad = [X'\Sigma^{-1}X + V^{-1}]^{-1}[X'\Sigma^{-1}Y + V^{-1}b]$$

and posterior covariance

$$V_Y = \mathsf{E}[(\beta - b_Y)(\beta - b_Y)' \mid Y] = [X'\Sigma^{-1}X + V^{-1}]^{-1}.$$

In the "noninformative" limit as $V^{-1} \to 0$, $b_Y \to \hat{\beta}$ and V_Y converges to the covariance matrix for $\hat{\beta}$, so the frequentist and Bayesian estimates and error bounds coincide in the limit as prior information becomes more and more diffuse or the data more copious.

Both prior and posterior distributions for β are p-variate normal, but with different mean and covariance parameters:

$$\text{Prior: } \beta \sim \mathsf{No}(b, V) \qquad \text{Posterior: } \beta \mid Y \sim \mathsf{No}(b_Y, V_Y).$$

For the prior distribution, the smallest volume sets in $\mathbb{R}^p$ with a specified probability of containing β are ellipsoids of the form

$$\mathcal{E}(c) \equiv \left\{\beta : \ (\beta - b)' V^{-1} (\beta - b) \leq c\right\}$$

for some $c > 0$. Since $(\beta - b)' V^{-1} (\beta - b)$ has the χ^2_p distribution, such a set will contain β with probability γ (say, 90%) if we take $c = \chi^2_p(\gamma)$, the $100\gamma\%$ quantile of the χ^2 distribution with p degrees of freedom. The volume in $\mathbb{R}^p$ of such an ellipsoid is

$$\text{vol}\big(\mathcal{E}(c)\big) = \omega_p \ \sqrt{\det(cV)}$$

where $\omega_p \equiv \pi^{p/2}/\Gamma(1 + p/2)$ is the volume of the unit ball in $\mathbb{R}^p$. Thus for any $0 < \gamma < 1$ the *posterior* $100\gamma\%$ HPD set is smaller in volume than the *prior* $100\gamma\%$ HPD by a factor of

$$\begin{aligned}
\frac{\omega_p \ \sqrt{\det(\chi^2_p(\gamma) V)}}{\omega_p \ \sqrt{\det(\chi^2_p(\gamma) V_Y)}} &= \frac{\det(V)^{1/2}}{\det\left(X'\Sigma^{-1}X + V^{-1}\right)^{-1/2}} \\
&= \left\{\det(V) \det\left(X'\Sigma^{-1}X + V^{-1}\right)\right\}^{1/2} \\
&= \left\{\det\left(I + V^{1/2} X'\Sigma^{-1} X V^{1/2}\right)\right\}^{1/2}.
\end{aligned}$$

For isotropic prior covariance $V = v^2 I$ and eigenvalues $\lambda_1 \geq ... \geq \lambda_p \geq 0$ for $X'\Sigma^{-1}X$, this is

$$\text{Volume shrinkage ratio } = \left\{\prod_{i=1}^{p} (1 + v^2 \lambda_i)\right\}^{1/2} \geq v^r \prod_{i=1}^{r} \sqrt{\lambda_i}, \tag{3}$$

bounded below for any $r \in \{1, \cdots, p\}$ by v^r times the square root of the product of the first r principal values.

References

Novak, J., Novak, K., Pratt, S., Vredevoogd, J., Coleman-Smith, C. E., and Wolpert, R. L. (2012), "Determining Fundamental Properties of Matter Created in Ultrarelativistic Heavy-Ion Collisions," Preprint, Department of Physics and Astronomy, Michigan State University, Draft 2012-11-26.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–435.

Last edited: March 20, 2014