Part III*

1. SS-ANOVA Spaces on General Domains

2. Averaging Operators and ANOVA Decompositions

3. Reproducing Kernel Spaces for ANOVA Decompositions

4. Building Blocks for SS-ANOVA Spaces, General and Particular

5. Representation of SS-ANOVA Fits

6. Example: Risk of Progression of Diabetic Retinopathy in the WESDR Study. Bernoulli data.

*Part III of 'An Introduction to Model Building With Reproducing Kernel Hilbert Spaces', by Grace Wahba, Univ. of Wisconsin Statistics Department TR 1020, Overheads for Interface 2000 Short Course.

7. GACV for smoothing parameters in the Bernoulli case.

• General Model Domains: $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$.
SS-ANOVA spaces.

Let $t = (t_1, \cdots, t_d),\ \ t_\alpha \in \mathcal{T}^{(\alpha)}, \alpha = 1, \cdots d$. Some examples are:

$$
\begin{array}{rcl}
\mathcal{T}^{(\alpha)} & = & [0,1] \quad unit\ \ interval \\
\mathcal{T}^{(\alpha)} & = & E^r \quad Euclidean\ \ r-space \\
\mathcal{T}^{(\alpha)} & = & \mathcal{S} \quad the\ \ sphere \\
\mathcal{T}^{(\alpha)} & = & \{1, \cdots, N\} \quad ordered\ \ categorical \\
\mathcal{T}^{(\alpha)} & = & \{\diamondsuit, \triangle, \heartsuit\} \quad unordered\ \ categorical \\
\cdots & & \cdots
\end{array}
$$

We let $t \in \mathcal{T} \equiv \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$. Let $\mathcal{E}$ be an averaging operator on $\mathcal{T}^{(\alpha)}$, defined by

$$\mathcal{E}_\alpha f_\alpha = \int_{\mathcal{T}^{(\alpha)}} f_\alpha d\mu_\alpha$$

where $d\mu_\alpha$ is some given probability distribution on $\mathcal{T}^{(\alpha)}$, for example if $\mathcal{T}^{(\alpha)} = [0,1]$ the uniform distribution is convenient. Given the $\mathcal{E}_\alpha$, any real valued function $f(t) = f(t_1, \cdots, t_d)$ on $\mathcal{T}$ has an ANOVA decomposition as follows:

Here's the ANOVA decomposition:

$$\begin{aligned}
f(t) &= \mu + \sum_{\alpha} f_{\alpha}(t_{\alpha}) \\
&+ \sum_{\alpha \leq \beta} f_{\alpha\beta}(t_{\alpha}, t_{\beta}) \\
&+ \cdots + f_{1,\cdots,d}(t_1, \cdots, t_d)
\end{aligned}$$

where the components are generated by the decomposition of the identity:

$$\begin{aligned}
f &= \prod_{\alpha} [\mathcal{E}_{\alpha} + (I - \mathcal{E}_{\alpha})] f \\
f &= \prod_{\alpha} \mathcal{E}_{\alpha} f + \sum_{\alpha} (I - \mathcal{E}_{\alpha}) \prod_{\beta \neq \alpha} \mathcal{E}_{\beta} f \\
&+ \sum_{\alpha < \beta} (I - \mathcal{E}_{\alpha})(I - \mathcal{E}_{\beta}) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_{\gamma} f \\
&+ \cdots + \prod_{\alpha = 1}^{d} (I - \mathcal{E}_{\alpha}) f.
\end{aligned}$$

(from the previous slide)

$$\begin{aligned} f &= \prod_\alpha \mathcal{E}_\alpha f + \sum_\alpha (I-\mathcal{E}_\alpha) \prod_{\beta\neq\alpha} \mathcal{E}_\beta f \\ &\quad + \sum_{\alpha<\beta} (I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta) \prod_{\gamma\neq\alpha,\beta} \mathcal{E}_\gamma f \\ &\quad + \cdots + \prod_{\alpha=1}^{d} (I-\mathcal{E}_\alpha) f. \end{aligned}$$

Therefore:

$$\begin{aligned} \mu &= \prod_\alpha \mathcal{E}_\alpha f, \quad f_\alpha = (I-\mathcal{E}_\alpha) \prod_{\beta\neq\alpha} \mathcal{E}_\beta f \\ f_{\alpha,\beta} &= (I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta) \prod_{\gamma\neq\alpha,\beta} \mathcal{E}_\gamma f \\ \cdots & \quad \cdots \\ f_{1,2,\cdots,d} &= \prod_{\alpha=1}^{d} (I-\mathcal{E}_\alpha) f, \end{aligned}$$

and satisfy the `ANOVA SIDE CONDITIONS`

$$\begin{aligned} \mathcal{E}_\alpha f_\alpha &= 0 \\ \mathcal{E}_\alpha f_{\alpha\beta} &= \mathcal{E}_\beta f_{\alpha\beta} = 0 \\ \mathcal{E}_\alpha f_{\alpha\beta\gamma} &= \mathcal{E}_\beta f_{\alpha\beta\gamma} = \mathcal{E}_\gamma f_{\alpha\beta\gamma} = 0 \\ &\vdots \end{aligned}$$

Now, let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ for all $f_\alpha(t_\alpha) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. We can construct an RKHS $\mathcal{H}$ as the direct sum of subspaces which correspond to this decomposition:

$$
\begin{aligned}
\mathcal{H} &= \prod_{\alpha=1}^{d} [\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}] \\
\mathcal{H} &= [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \\
&\oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \\
&\oplus \cdots \oplus \prod_{\alpha=1}^{d} \otimes \mathcal{H}^{(\alpha)}.
\end{aligned}
$$

($[1]^{(\gamma)}$ are omitted wherever they occur).

Let $R_\alpha(s_\alpha, t_\alpha)$ be the RK for $\mathcal{H}^{(\alpha)}$. A (Smoothing Spline) ANOVA space $\mathcal{H}_K$ of functions on $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$ is given by:

$$\mathcal{H}_K = \prod_{\alpha=1}^{d} [[1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)}]$$

which then has the RK

$$\begin{aligned} K(s,t) &= \prod_{\alpha=1}^{d} [1 + R_\alpha(s_\alpha, t_\alpha)] \\ &= 1 + \sum_{\alpha=1}^{d} R_\alpha(s_\alpha, t_\alpha) \\ &\quad + \sum_{\alpha<\beta} R_\alpha(s_\alpha, t_\alpha) R_\beta(s_\beta, t_\beta) \\ &\quad + \cdots + \prod_{\alpha=1}^{d} \otimes R_\alpha(s_\alpha, t_\alpha). \end{aligned}$$

• SS-ANOVA spaces, continued.

$\mathcal{H}^{(\alpha)}$ may be further decomposed into a low dimensional parametric part $\mathcal{H}_\pi^{(\alpha)}$, and a 'smooth' part $\mathcal{H}_s^{(\alpha)}$, $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_s^{(\alpha)}$. This can be done in many ways, depending on the $\mathcal{T}^{(\alpha)}$ and what part of the model it is desired not to penalize. A useful example when $\mathcal{T}^{(\alpha)} = [0, 1]$, which will be employed later is:

$$\begin{array}{rcccl} \mathcal{H}^{(\alpha)} & = & \{k_1\} & \oplus & [\{k_2\} \oplus W_2^0] \\ R_\alpha(s_\alpha, t_\alpha) & = & r_\pi(s_\alpha, t_\alpha) & + & r_s(s_\alpha, t_\alpha) \quad say \end{array}$$

where

$$\begin{aligned} r_\pi(s_\alpha, t_\alpha) &= k_1(s_\alpha)k_1(t_\alpha), \\ r_s(s_\alpha, t_\alpha) &= k_2(s_\alpha)(k_2(t_\alpha)) - k_4([s_\alpha - t_\alpha]) \end{aligned}$$

We have encountered r_s before, the square norm in its associated RKHS is $\int_0^1 (f'')^2$. In this example the $\mathcal{T}^{(\alpha)}$ and r_π and r_s will be the same for each component of t, but this not necessary.

• SS-ANOVA spaces, continued.

Now $K(s,t)$ can be seen to be expandable in the tensor sums and products of $r_\pi(s_\alpha, t_\alpha)$ and $r_s(s_\alpha, t_\alpha)$, $\alpha = 1, \cdots, d$. The expansion is carried out and truncated (Model selection!), in our experience, interactions higher than two-factor can generally be deleted, and frequently only a few two factor interactions are important. Finally, terms containing only r_π's will not be penalized, and are collected into $\mathcal{H}_0$, and the spanning set for $\mathcal{H}_0$ will be relabeled as $\{\phi_1, \cdots, \phi_M\}$, The terms with one or more r_s are collected into $\mathcal{H}_1$, and relabeled as $\mathcal{H}_1 = \sum_\beta \mathcal{H}^\beta$, with the RK's Q^β for the $\mathcal{H}^\beta$ weighted and relabeled as

$$Q_\theta(s,t) = \sum_\beta \theta_\beta Q^\beta(s,t).$$

Note that the Q^β generally depend only on a subset of the components of (s,t). The θ_β allow for different smoothing parameters for the different components.

• SS-ANOVA spaces, continued.

We have, finally, reduced an arbitrary ANOVA model to the case established via the representer theorem: Find $f = f_0 + f_1$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$ to min

$$I_\lambda(y, f) = \frac{1}{n} \sum_{i=1}^{n} g_i(y_i, f(t(i))) + \lambda \sum_\beta \theta_\beta^{-1} \|P^\beta f\|^2_{\mathcal{H}_{Q_\beta}}.$$

where $P^\beta f$ is the component of f in $\mathcal{H}_{Q_\beta}$. Then the minimizer f_λ of I_λ is unique and has, by the representer theorem, the representation

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i Q_\theta(t(i), \cdot).$$

• SS-ANOVA Example: Risk of Progression of Diabetic Retinopathy in the Younger Onset Population in the Wisconsin Epidemiolocig Study of Diabetic Retinopathy.

Data:

$$\{y_i, t(i)\}, t = (\texttt{dur}, \texttt{gly}, \texttt{bmi}), i = 1, \cdots, n = 669.$$

where

$$\begin{aligned} y_i &= 1, \ \ progression\ yes \\ &= 0, \ \ progression\ no \\ \texttt{dur} &= \ \ duration\ of\ diabetes\ at\ baseline \\ \texttt{gly} &= \ \ glycosylated\ hemoglobin \\ \texttt{bmi} &= \ \ body\ mass\ index \end{aligned}$$

Goal: Estimate $p(t)$, the probability of progression given t. Let $f(t) = log[p(t)/(1-p(t))]$. The $-loglik(y, f)$ is

$$-log[p^y(1-p)^{1-y}] \equiv -yf + log(1 + e^f) \equiv g(y, f)$$

• SS-ANOVA Example: Diabetic Retinopathy (con't).

We selected the model

$$\begin{aligned} f(\mathtt{dur}, \mathtt{gly}, \mathtt{bmi}) &= \mu + f_1(\mathtt{dur}) + a_2 \cdot \mathtt{gly} \\ &+ f_3(\mathtt{bmi}) + f_{13}(\mathtt{dur}, \mathtt{bmi}) \end{aligned}$$

$$\mathcal{H}_0 :$$

$$\{\phi_\nu(\mathtt{dur}, \mathtt{gly}, \mathtt{bmi})\} = \{1, k_1(\mathtt{dur}), k_1(\mathtt{gly}), k_1(\mathtt{bmi})\}$$

$$\mathcal{H}_1 :$$

β	$Q^\beta(\mathtt{dur}, \mathtt{bmi}; \mathtt{dur}', \mathtt{bmi}')$
1	$r_s(\mathtt{dur}, \mathtt{dur}')$
2	$r_s(\mathtt{bmi}, \mathtt{bmi}')$
3	$r_\pi(\mathtt{dur}, \mathtt{dur}') r_s(\mathtt{bmi}, \mathtt{bmi}')$
4	$r_s(\mathtt{dur}, \mathtt{dur}') r_\pi(\mathtt{bmi}, \mathtt{bmi}')$
5	$r_s(\mathtt{dur}, \mathtt{dur}') r_s(\mathtt{bmi}, \mathtt{bmi}')$

$$Q_\theta(\mathtt{dur}, \mathtt{bmi}; \cdot)) = \sum_{\beta=1}^{5} \theta_\beta Q^\beta(\mathtt{dur}, \mathtt{bmi}; \cdot).$$

●SS-ANOVA EXAMPLE: Diabetic Retinopathy (con't).

We will minimize

$$I_\lambda(y,f) = \frac{1}{n}\sum_{i=1}^{n}[-loglik(y_i,f_i) + \lambda\sum_{\beta}\theta_\beta^{-1}\|P^\beta f\|^2_{\mathcal{H}_{Q_\beta}}.$$

where $f_i = f(t(i))$, $P^\beta f$ is the component of f in $\mathcal{H}_{Q_\beta}$. The minimizer f_λ of I_λ has the representation

$$f(\cdot) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(\cdot) + \sum_{i=1}^{n} c_i Q_\theta(t(i),\cdot),$$

and we need to compute (d,c) to min

$$\frac{1}{n}\sum_{i=1}^{n}(-y_i f_i + log(1+e^{f_i})) + \lambda c' K c$$

where $f_i = (Td+Kc)_i, T_{n\times 4} = \{\phi_\nu(t(i))\}, K_{n\times n} = Q_\theta(t(i),t(j))$; and we need to estimate $\lambda_\beta = \lambda\theta_\beta, \beta = 1,\cdots,5$.

♣♣ Choosing $\lambda = (\lambda_1, \cdots, \lambda_q)$, Bernoulli data.

Notation: Let $f_\lambda^{[k]}(\cdot)$ be the minimizer of $I_\lambda(y, f)$ with the kth data point omitted. Let $f_{\lambda k}^{[k]} = f_\lambda^{[k]}(t(k))$, $f_{\lambda k} = f_\lambda(t(k))$. Let $b(f) = log(1 + e^f)$, thus $g(y, f) = -yf + b(f)$.

• Leaving-out-one.

Choose λ to min

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^{n} [-y_k f_{\lambda k}^{[k]} + b(f_{\lambda k})].$$

Generally not practical.

• GACV (Generalized Approximate Cross Validation).

The inverse Hessian $H(\lambda)$ of I_λ with respect to $(f_{\lambda 1}, \cdots, f_{\lambda n})$ at the minimizer plays an important role. It is an interesting fact that the influence matrix $A(\lambda)$ in the Gaussian case is also the inverse Hessian of I_λ in the Gaussian case, and this is true to first order in the general exponential family case, and in some other situations. H can be thought of as the (local) influence matrix, since in the nonquadratic case it depends on f_λ. It can be shown that

$$V_0(\lambda) \approx ACV(\lambda) = \frac{1}{n}\sum_{k=1}^{n}[-y_k f_{\lambda k} + b(f_{\lambda k})] + D_0(\lambda).$$

where

$$D_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{h_{ii} y_i (y_i - p_{\lambda i})}{[1 - h_{ii}\sigma_{ii}]},$$

h_{ii} is the iith entry of $H(\lambda)$, $p_{\lambda i} = e^{f_{\lambda i}}/(1 + e^{f_{\lambda i}})$, $\sigma_{ii} = p_{\lambda i}(1 - p_{\lambda i})$. The GACV is obtained from the ACV by replacing h_{ii} and $h_{ii}\sigma_{ii}$ by their averages, as follows:

In the expression for D_0 h_{ii} is replaced by $\frac{1}{n}\sum_{i=1}^n h_{ii} \equiv \frac{1}{n}tr(H)$ and $1-h_{ii}\sigma_{ii}$ is replaced by $\frac{1}{n}tr[I-(W^{1/2}HW^{1/2})]$, where $W = diag\{\sigma_{ii}\}$, giving

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + b(f_{\lambda i})]$$
$$+\frac{tr(H)}{n}\frac{\sum_{i=1}^n y_i(y_i - p_{\lambda i})}{tr[I-(W^{1/2}HW^{1/2})]}.$$

The randomized trace technique may be used to evaluate $GACV$:

$$ranGACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{\lambda i} + b(f_{\lambda i})]$$
$$+\frac{\delta'(f_\lambda^{y+\delta} - f_\lambda^y)}{n}\frac{\sum_{i=1}^n y_i(y_i - p_{\lambda i})}{[\delta'\delta - \delta' W(f_\lambda^{y+\delta} - f_\lambda^y)]}.$$

δ is a random white noise perturbation n-vector and $f_\lambda^{y+\delta}$ is the n- vector of values of the fit at the observation points **based on estimating f with perturbed data** $y+\delta$. We show next that

$$\delta'(f_\lambda^{y+\delta} - f_\lambda^y)$$

provides a randomized estimate of $traceH(\lambda)$.

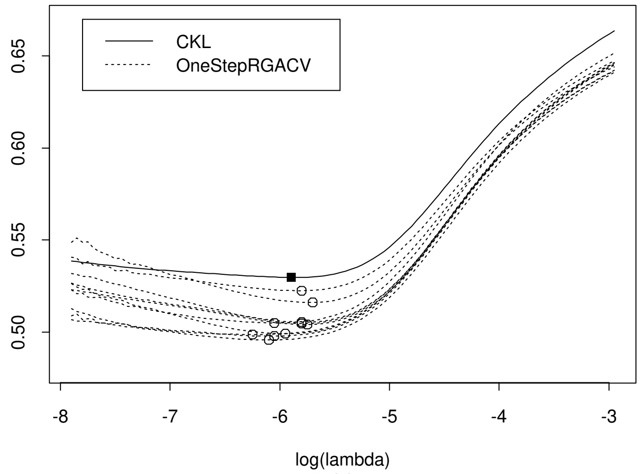
• Randomized Trace Estimates.

δ is a (small) random perturbation with $E\delta = 0$ and $cov\delta = \sigma_\delta I$. For any matrix H, $E\delta' H\delta = \sigma_\delta traceH$. Now, let $H[\cdot]$ be the operator which maps a data vector z into the vector of values of f_λ at the observation points, that is, $H[z] = f_\lambda^z$. In the Gaussian case H is linear and we just have $H[z] = Hz$. We have, to first order

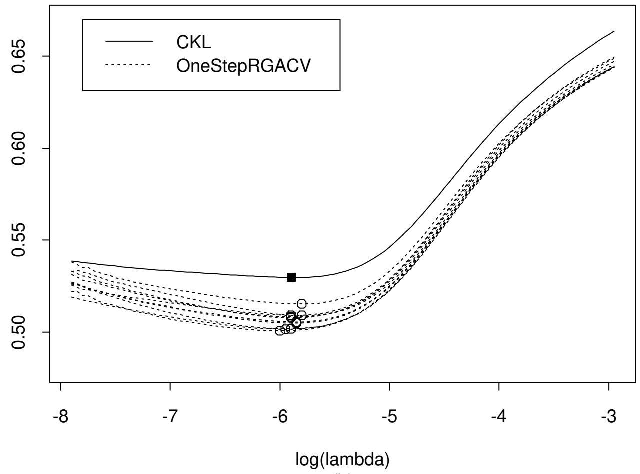
$$f_\lambda^{y+\delta} - f_\lambda^y \approx H[y+\delta] - H[y] \approx H[y*]\delta$$

where $y*$ is some intermediate value between $y + \delta$ and y. Thus, we have the approximation

$$E\delta^{'}(f_\lambda^{y+\delta} - f_\lambda^y) \sim \delta' H[y*]\delta \approx \sigma_\delta trH[y)]$$



(a)



(b)

Figure 5a. 10 replicates of $ranGACV(\lambda)$, compared to $CKL(\lambda)$, where the Comparative Kullback-Liebler distance (CKL) is given by $CKL(\lambda)[p_\lambda, p_{TRUE}] = \frac{1}{n}\sum_{i=1}^{n}[-p_{TRUEi}f_{\lambda i} + b(f_{\lambda i})]$

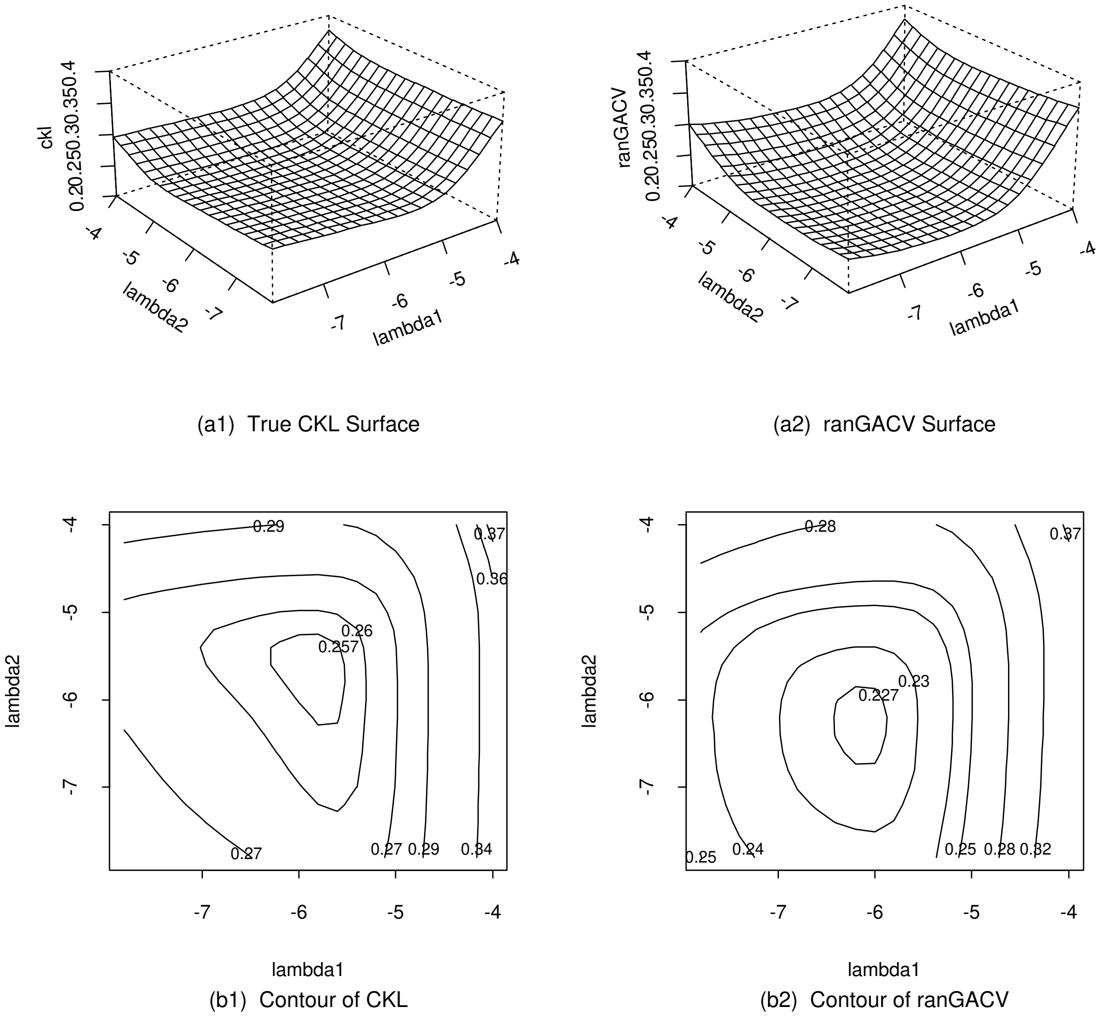


Figure 5b. $ranGACV$, compared to the true CKL, $\lambda = (\lambda_1, \lambda_2)$. Left: CKL. Right: $ranGACV$.

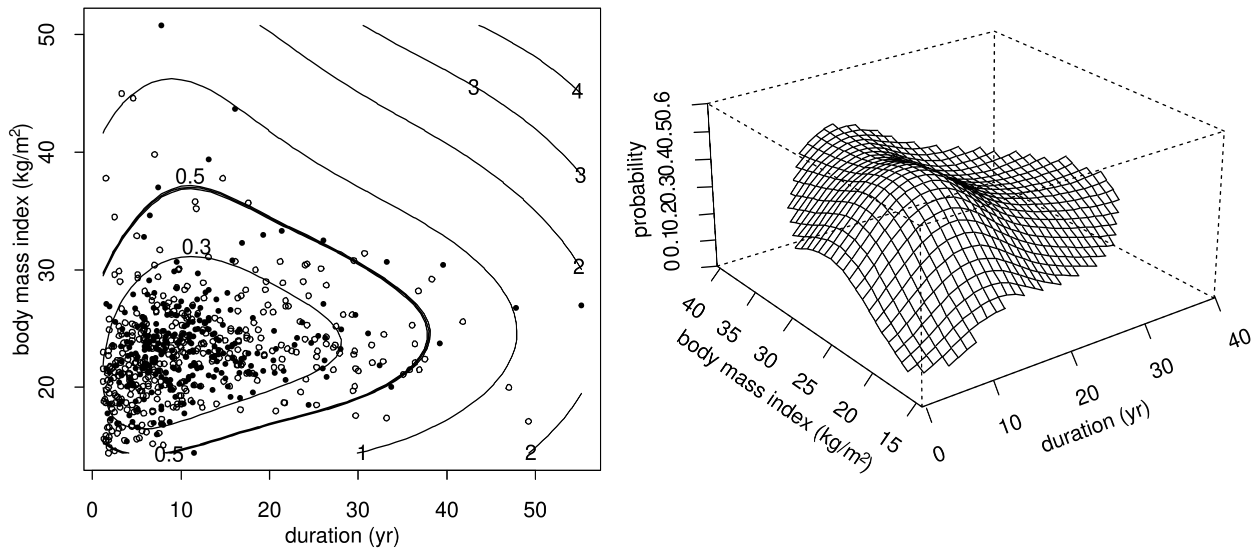


Figure 6a. Left: Data and contours of constant posterior standard deviation. Right: Estimated probability of progression as a function of duration and body mass index for glycosylated hemoglobin fixed at its median.

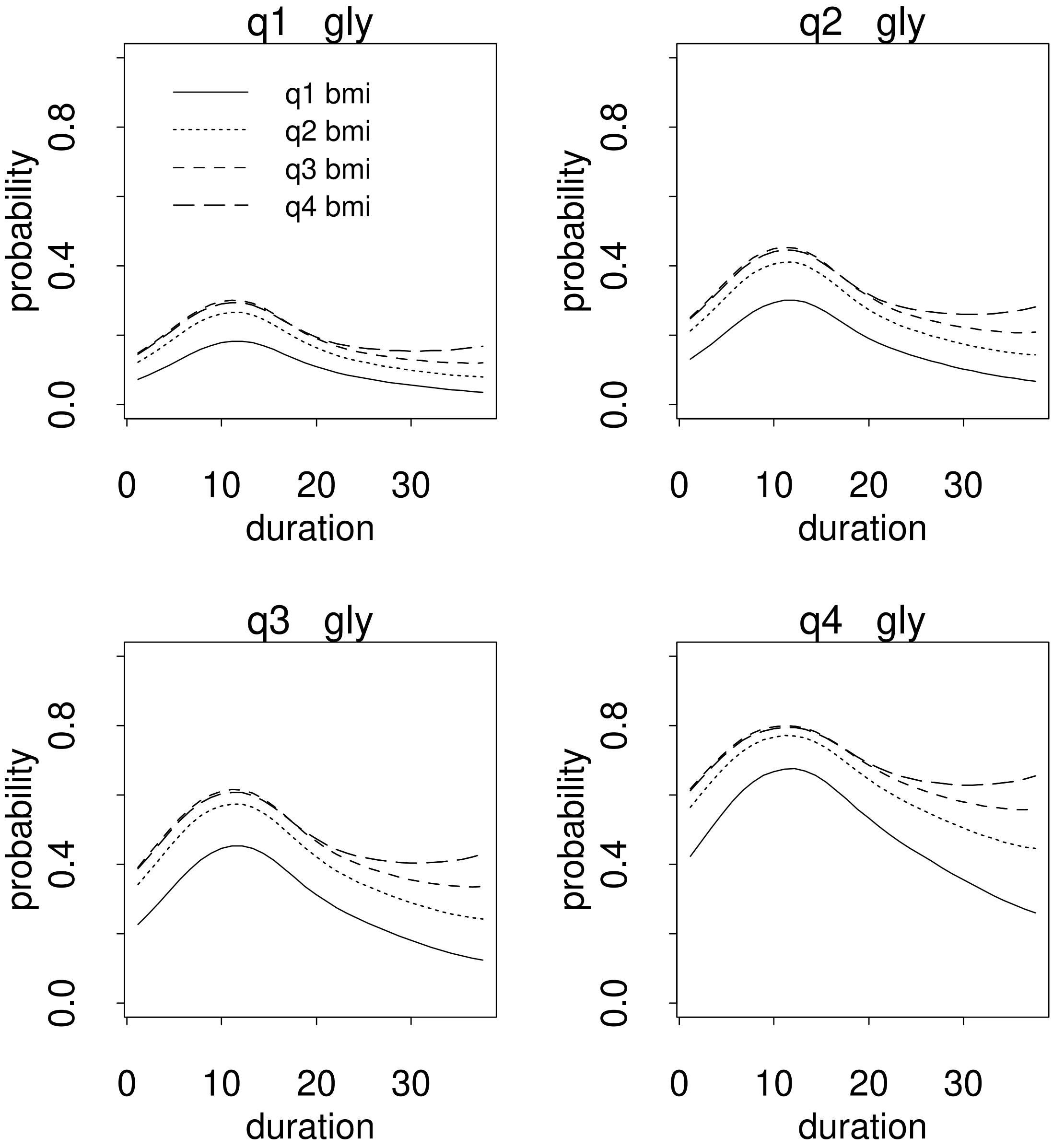


Figure 6b. Estimated probability of progression as a function of `dur` for four levels of `bmi` by four levels of `gly`.

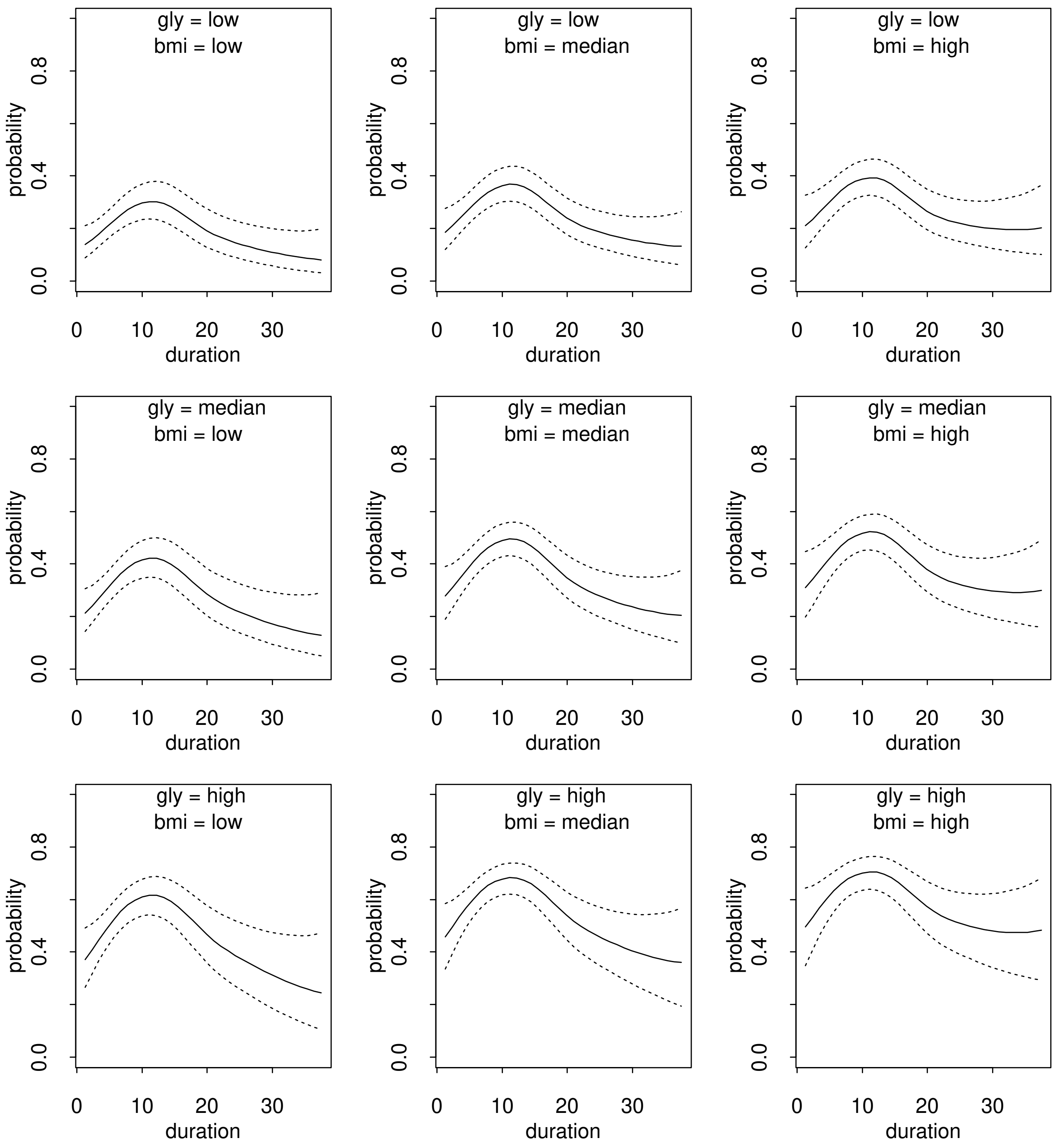


Figure 6c. Bayesian ‘Confidence Intervals’

• SS-ANOVA spaces, SOFTWARE.

Codes for SS-ANOVA models, reverse chronological order: Use a Newton-Raphson algorithm for (d, c) given λ. Use an iterative unbiased risk estimate for λ in the Bernoulli case.

...

Code- Author- Where Found $(* = freeware)$

—- —– —–

$*$ gss- Chong Gu- `http:www.r-project.org`

$*$ GRKPACK-Yuedong Wang-`http:www.netlib.org/gcv`

$*$ RKPACK- Chong Gu-`http:www.netlib.org/gcv` recap

Part I

1. Positive Definite Functions

2. Bayes Estimates and Variational Problems

3. Reproducing Kernel Hilbert Spaces

4. The Moore-Aronszajn Theorem and Inner Products in RKHS

5. Example: Periodic Splines

6. The Representer Theorem (simple case)

7. Sums and Products of Positive Definite Functions

Part II

1. The polynomial smoothing spline.

2. Leaving-out-one, GCV and other smoothing parameter estimates.

3. The thin plate smoothing spline.

4. Generalizations: Different kinds of observations: Non-gaussian, indirect, constrained.

5. Examples: The histospline, convolution equations with positivity constraints. GCV with inequality constraints.

Part III

1. SS-ANOVA Spaces on General Domains
2. Averaging Operators and ANOVA Decompositions
3. Reproducing Kernel Spaces for ANOVA Decompositions
4. Building Blocks for SS-ANOVA Spaces, General and Particular
5. Representation of SS-ANOVA Fits
6. Example: Risk of Progression of Diabetic Retinopathy in the WESDR Study. Bernoulli data.
7. GACV for smoothing parameters in the Bernoulli case.

♣♣ Ending Comments

Reproducing Kernel Hilbert Spaces apparently first appeared in the Statistics Literature in the work of Parzen in the late 60's, and although there was theoretical work in the early 70's there were several things that were necessary to make models based on them useful to the data analyst: (i) high speed computers that could handle the solution of large linear systems, (ii) method(s) for choosing the smoothing parameter(s), (iii) user friendly software, since the models are generally non-trivial to code from scratch. These things have come to pass for some models, but for some of the more recent methods, user-friendly software is not (yet) available. There are still many interesting open theoretical and practical problems for the research-minded - particularly related to variable and model selection in very large, complex data sets, and efficient code development. However, we hope we have shown that model building with RKHS has the flexibility and generality to handle a very wide variety of statistical data analysis problems, and have given the interested user ideas on how to begin doing this.