Penalized Functional Regression for Next-Generation Sequencing Studies

Olga A. Vsevolozhskaya, Michigan State University

Available at: http://works.bepress.com/vsevolozhskaya/20/
Penalized Functional Regression for Next-Generation Sequencing Studies

Olga Vsevolozhskaya

Michigan State University

August 11, 2014
Figure: Genetic data of a single individual over 20 different SNPs.
Genotype Function

Figure: Genotype function $G_i(t)$; $t$ represents SNP position in the genetic region.

Luo et al. (2012): $y_i = \beta_0 + \int_0^1 G_i(t)\beta(t)dt + \epsilon_i$, $\epsilon_i \text{i.i.d.} \sim N(0, \sigma^2)$. 
Vsevolozhskaya et al. (2014): $G_{ij}(t) = \mu_i(t) + \epsilon_{ij}(t)$, $\epsilon_{ij}(t) \overset{i.i.d.}{\sim} \text{GP}(0, \gamma)$.

**Figure**: Genotype functions of cases versus controls.
### Flipping Method

<table>
<thead>
<tr>
<th>SNP₁</th>
<th>SNP₂</th>
<th>Flip?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>✓</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>✗</td>
</tr>
</tbody>
</table>

Sum = 3

<table>
<thead>
<tr>
<th>SNP₁</th>
<th>SNP₂</th>
<th>Flip?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>✗</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>✗</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>✗</td>
</tr>
</tbody>
</table>

Sum = 0

**Table**: Sample genotype information over seven subjects and two SNPs. *Left panel* illustrates original data and highlights subjects for which a flip in genotype coding at SNP₂ position is possible. *Right panel* illustrates “flipped” data at the SNP₂ position.
Flipping Method

Figure: Smooth genotype curves using original data (right panel), and relabeled data (left panel).
Suppose we have 2 independent groups with functional samples $G_{i1}(t), \ldots, G_{in_i}(t), \ i = 1, 2$, $t \in \mathcal{T}$ denotes SNP position $\mathcal{T}$ region.

We wish to test if the mean genotype functions vary between two groups over a continuous sequence region $\mathcal{T}$:

$$H_0 : \mu_1(t) = \mu_2(t)$$

$$H_a : \mu_1(t) \neq \mu_2(t)$$

The test statistics:

$$F = \frac{\int_{\mathcal{T}} \sum_{i=1}^{k} n_i (\hat{\mu}_i(t) - \hat{\mu}(t))^2 dt / (k - 1)}{\int_{\mathcal{T}} \hat{\gamma}(t, t) dt},$$

where $\hat{\mu}_i(t)$'s are the group mean functions and $\hat{\gamma}(t, t)$ is the common covariance function.
Penalized Sums of Squares

• **How do we estimate the group mean functions, \( \hat{\mu}_i(t) \)'s?**

• Consider a simpler problem of estimating a single curve from the observed \( y_i \)'s and \( t_i \)'s:

\[
y_i = f(t_i) + \epsilon_i, \quad \epsilon_j \sim N(0, \sigma^2) \quad i = 1, \ldots, T \quad t_i \in [0, 1]
\]

• One possible approach is to minimize penalized least squares.

\[
\hat{f}(t_i) = \min \left\{ n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^1 [Ly(t)]^2 \, dt \right\}.
\]

Minimizing just the mean squared error may result in data overfit or in a curve with a high sampling variance, manifested in the rapid local variation of the curve.

**Figure**: Minimizing just MSE.
Penalized Sums of Squares

• **How do we estimate the group mean functions, \( \hat{\mu}_i(t) \)'s?**
• Consider a simpler problem of estimating a single curve from the observed \( y_i \)'s and \( t_i \)'s:
  \[
y_i = f(t_i) + \epsilon_i, \quad \epsilon_j \sim N(0, \sigma^2) \quad i = 1, \ldots, T \quad t_i \in [0, 1]
  \]
• One possible approach is to minimize penalized least squares.

\[
\hat{f}(t_i) = \min \left\{ n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_{0}^{1} [L y(t)]^2 \, dt \right\}.
\]

  • The “roughness” of the function can be quantified by a penalty term.
  • \( Ly(t) \) is a linear differential operator.

For example:
\[
Ly(t) = \beta y(t) + D y(t) = \beta y(t) + \frac{dy}{dt} = 0,
\]

  corresponding to the 1st order homogeneous differential equation:
  \[
y' = \beta y.
\]

Figure: Functional fit with additional roughness penalty.
• **How do we choose roughness penalty** $L_y$?

• The square of the second derivative $[D^2y(s)]^2 ds$ is a common choice, so the penalized sums of squares equation becomes:

$$\hat{f}(t_i) = \min \{ n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_{0}^{1} [y''(t)]^2 dt \}. \quad (1)$$

• $[D^2y(s)]^2 ds$ is often called a *curvature* of a function at $t$.

• Note: A straight line $b + kt$ has no curvatures and will not be penalized by the square of the second derivative.

• The smoothing parameter should be specified by a used or determined through GCV.
Penalized Sums of Squares

• **How do we choose roughness penalty** $L_y$?

• The square of the second derivative $[D^2y(s)]^2 ds$ is a common choice, so the penalized sums of squares equation becomes:

\[
\hat{f}(t_i) = \min \left\{ n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^1 [y''(t)]^2 \, dt \right\}.
\]

• $[D^2y(s)]^2 ds$ is often called a *curvature* of a function at $t$.

• Note: A straight line $b + kt$ has no curvatures and will not be penalized by the square of the second derivative.

• The smoothing parameter should be specified by a user or determined through GCV.
BLUP and Penalized Sums of Squares

- Kimeldorf and Wahba (1971) showed that for any $\lambda > 0$, the function $f$ minimizing the spline smoothing criterion defined by a linear differential operator $L$ of order $m$ has the expansion:

$$\hat{f}(t) = \sum_{j=1}^{m} d_j \phi_j(t) + \sum_{i=1}^{n} c_i k_2(t_i, t).$$  \hspace{1cm} (2)

- Terry Speed (Speed (1991)) in his discussion of Geoff Robinson’s original paper (Robinson (1991)) noted that this solution to the minimization problem is a case of BLUP.

$$\hat{f}(t) = X\hat{d} + Z\hat{c},$$

- The matrices $X$ and $Z$ are user-supplied components and are defined by the particular choice of operator $L$ used in the smoothing application.
How do we specify $X$ and $Z$?

1. $\phi_j, j = 1, \ldots, m$: a set of linearly independent function satisfying $L\phi_j = 0$, that is, the basis of the nulls space, $\ker L$.
   - If $[D^2 y(s)]^2 ds$ is used to measure roughness, then $\phi_1 = 1$ and $\phi_2 = t$ are the two fundamental solutions to $y'' = 0$ and span the null space of $L$.

2. $k_2$: the reproducing kernel function that is found as $k_2(t, s) = \int G(s; w)G(t; w)dw$, where $G$ is the Green’s function for a linear differential operator $L$.
   - Green’s function corresponding to the homogeneous ODE $y'' = 0$ is $G(s; w) = (s - w)_+$, where $(x)_+ = x$ for all $x \geq 0$ and $(x)_+ = 0$ otherwise.
Back to the model:

\[ y_i = f(t_i) + \epsilon_i, \quad \epsilon_j \sim N(0, \sigma^2) \quad i = 1, \ldots, T \quad t_i \in [0, 1]. \]

To find the minimizer of

\[ \hat{f}(t_i) = \min \left\{ n^{-1} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^1 [y''(t)]^2 \, dt \right\}, \]

for a sequence of knots \( \kappa_1, \ldots, \kappa_K \), construct

\[
X = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \quad Z = \begin{bmatrix} (t_1 - \kappa_1)_+ & \ldots & (t_1 - \kappa_K)_+ \\ \vdots & \vdots & \vdots \\ (t_n - \kappa_1)_+ & \ldots & (t_n - \kappa_K)_+ \end{bmatrix}
\]

and fit \( \hat{f}(t) = d_1 + d_2 t + \sum_{i=1}^{K} (t - \kappa_i)_+ c_i \) using standard fixed effect software.

The corresponding smoothing parameter will be \( \lambda = \sigma^2_{\epsilon} / \sigma^2_c \).
Once the design matrices $X$ and $Z$ are defined, the model can be fit in R as:

```r
require(nlme)
group <- rep(1, N) # N is the number of subjects
fit.lme <- lme(y~-1+X, random=list(group=pdIdent(~-1+Z)))

require(mgcv)
fit.gam. <- gam(y~-1+X+Z,
    paraPen=list(Z=list(diag(Kb-2))),
    method="REML") # Kb is the number of knots
```

Additionally, **refund** package Ciprian Crainiceanu et al. (2013) can be used.
Consider a model with two predictors – one categorical (group effect) and one continuous:

\[ y_{ij}(t) = \mu(t) + \beta_1(t)X_{1i} + \beta_2(t)X_{2ij} + \epsilon_{ij}(t). \]

- \( X_{1i} \) takes 0/1 values; \( X_{2ij} \sim N(\mu, \sigma^2) \). Expand \( \mu(t), \beta_1(t), \beta_2(t) \) using \( \phi_i \)'s and \( k_{2j} \)'s as

\[ y_{ij}(t) = (b_0 + b_1 t + \sum k_2(t, t_i)b_i) + (b_0^* + b_1^* t + \sum k_2(t, t_i)b_i^*) \]

\[ + X_{2ij} (d_0 + d_1 t + \sum k_2(t, t_i)d_i). \]
FANOVA Test

- To test whether any of the effects is zero, $H_0 : \beta_i(t) = 0$, we can still use the function $F$ test statistic defined in Shen and Faraway (2004):

$$F = \frac{(rss_\omega - rss_\Omega)/(p - q)}{rss_\Omega/(n - p)} \approx \frac{\text{trace}(\hat{\Sigma}_\Omega - \hat{\Sigma}_\omega)/(p - q)}{\text{trace}(\hat{\Sigma}_\Omega)/(n - p)},$$

(3)

where $rss = \sum_{i=1}^{N} \int (y_i(t) - \hat{y}_i(t))^2 dt$, $\Omega$ represents a larger model with $\text{dim}(\Omega) = p$, $\omega$ is the reduced model with $\text{dim}(\omega) = q$, and $\hat{\Sigma}$ is the empirical covariance matrix.
Penalized Functional Regression for Next-Generation Sequencing Studies

Olga Vsevolozhskaya

References

Simulation Study

Figure: Simulations set up.

\[ A_i = \sum_j G_{i,j}^* \times \gamma_j + \epsilon_i, \quad i = 1, \ldots, \# \text{subjects}, \quad j = 1, \ldots, \# \text{variants}, \quad (4) \]

where * represents percentage of causal alleles, \( \gamma_j \sim N(\mu_\gamma, \sigma^2_\gamma) \) and \( \epsilon_j \sim N(0, 1) \).

\[ \text{logit}(\Pr(S_i = 1)) = \sum_j G_{i,j}^* \times \alpha_j + A_i \times \beta, \quad (5) \]

where \( \alpha_j \sim N(\mu_\alpha; \sigma_\alpha) \) and \( \beta \sim N(3, 1) \).
We considered three scenarios in our power simulations:

- the same set of risk variants for both phenotypes A and S (i.e., $G_{ij}^*$ is the same in Eq.(4) and Eq.(5));
- the same percentage but a random set of risk variants for A and S (i.e., the percentage of functional alleles $*$ is the same, but the causal SNPs are drawn independently for A and S);
- varying percentage of risk variants for A and a constant set of risk variants for S.
Penalized Functional Regression for Next-Generation Sequencing Studies

Olga Vsevolozhskaya

References

Power (Scenario 1)

Figure: Empirical power of FANOVA and SKAT for \( N = 50 \) subjects and the first simulation scenario. Panel (a): \( \mu_\gamma = \mu_\alpha = 0; \sigma_\gamma = \sigma_\alpha = 0.05 \). Panel (b): \( \mu_\gamma = \mu_\alpha = 0; \sigma_\gamma = \sigma_\alpha = 1 \). Panel (c): \( \mu_\gamma = \mu_\alpha = 0.05; \sigma_\gamma = \sigma_\alpha = 0.25 \). Panel (d): \( \mu_\gamma = \mu_\alpha = 0.05; \sigma_\gamma = \sigma_\alpha = 1 \).
Summary

- A natural way to deal with multiple phenotypes.
- No need to worry about a link function.
- Software for implementation is readily available.
Acknowledgments

• Dmitri V. Zaykin
• Qing Lu
• Jim C. Anthony
• Mark C. Greenwood


