A semivarying joint model for longitudinal binary and continuous outcomes

Supplementary Material

1 Estimation Procedure

We propose a two-stage estimation procedure. Before presenting the details of our procedure, we give a brief sketch of both stages. In the first stage we fit a semivarying coefficient model (Fan and Huang, 2005; Fan et al., 2007) to the continuous response. At this stage we employ the profile least squares approach proposed by Fan et al. (2007) to obtain efficient estimators of the regression coefficients $\boldsymbol{\alpha}_w(t)$ and $\boldsymbol{\beta}_w$. In the second stage we use the residuals from the first stage and the predictors for the binary response, and fit a generalized varying coefficient model for the binary response given the continuous response. At this stage we obtain the components necessary to compute the estimate of $\tau(t)$.

Now we turn to the details for the first stage. For a given $\boldsymbol{\beta}_w$, let $W_i^*(t) = W_i(t) - \mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_w$. Then the model for the continuous response becomes

$$W_i^*(t) = \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_w(t) + \varepsilon_{wi}(t), \tag{1}$$

which is a varying coefficient model (Cleveland et al., 1992; Hastie and Tibshirani, 1993). To estimate $\boldsymbol{\alpha}_w(t)$ in this model, we employ local linear fitting techniques (Fan and Gijbels, 1996), which leads to the following solution for the regression coefficients:

$$\hat{\boldsymbol{\alpha}}_w(t) = (\mathbf{I}_p, \mathbf{0}_p)(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\kappa}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\kappa}\mathbf{W}^*,$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\mathbf{0}_p$ is the $p \times p$ matrix of zeros, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_n)^{\mathrm{T}}$, $\boldsymbol{\Lambda}_i = ((1, t_{i1} - t) \otimes \mathbf{x}_{i1}, \ldots, (1, t_{in_i} - t) \otimes \mathbf{x}_{i\,n_i})$, and $\boldsymbol{\kappa}$ is an $N \times N$ diagonal matrix with the kernel weights along its diagonal.

Let $\mathbf{x} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}}, \mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{m} = (\mathbf{m}_1^{\mathrm{T}}, \ldots, \mathbf{m}_n^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{m}_i = (\mathbf{x}_i^{\mathrm{T}}(t_{i1})\boldsymbol{\alpha}_w(t_{i1}), \ldots, \mathbf{x}_i^{\mathrm{T}}(t_{i\,n_i})\boldsymbol{\alpha}_w(t_{i\,n_i}))^{\mathrm{T}}$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\varepsilon}_n^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_n(t_{in_i}))^{\mathrm{T}}$. Then (1) can be rewritten as

$$\mathbf{W} - \mathbf{z}\boldsymbol{\beta} = \mathbf{m} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{W} = \left(\mathbf{W}_1^{\mathrm{T}}, \ldots, \mathbf{W}_n^{\mathrm{T}}\right)^{\mathrm{T}}$ with $\mathbf{W}_i = (W_i(t_{i1}), \ldots, W_i(t_{in_i}))^{\mathrm{T}}$. This local linear regression produces an estimator linear in $W_i^*(t)$ (Fan and Gijbels, 1996), so that the estimate of $\boldsymbol{\alpha}_w(\cdot)$ is linear in $\mathbf{W} - \mathbf{z}\boldsymbol{\beta}$, and the estimator of $\mathbf{m}$ is $\hat{\mathbf{m}} = \mathbf{S}(\mathbf{W} - \mathbf{z}\boldsymbol{\beta})$. $\mathbf{S}$, which depends only on $\{t_{ij}, \mathbf{x}_i(t_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, n\}$, is usually referred to as the smoothing matrix of the local linear smoother.

Substituting $\hat{\mathbf{m}}$ in (2), we obtain

$$(\mathbf{I}_N - \mathbf{S})\mathbf{W} = (\mathbf{I}_N - \mathbf{S})\mathbf{z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

To estimate $\boldsymbol{\beta}_w$ more efficiently, we use weighted least squares, which yields

$$\hat{\boldsymbol{\beta}}_w = \{\mathbf{z}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{S})^{\mathrm{T}}\mathbf{R}(\mathbf{I}_N - \mathbf{S})\mathbf{z}\}^{-1}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{S})^{\mathrm{T}}\mathbf{R}(\mathbf{I}_N - \mathbf{S})\mathbf{W},$$

where $\hat{\boldsymbol{\beta}}_w$ is the profile weighted least squares estimator (Fan et al., 2007), $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{W} = \left(\mathbf{W}_1^{\mathrm{T}}, \ldots, \mathbf{W}_n^{\mathrm{T}}\right)^{\mathrm{T}}$ with $\mathbf{W}_i = (W_i(t_{i1}), \ldots, W_i(t_{in_i}))^{\mathrm{T}}$, $\mathbf{R}$ is the working covariance matrix, and $\mathbf{S}$ is the smoothing matrix of the local linear smoother. Misspecification of the working covariance matrix affects only the efficiency, not the consistency, of this estimator, whereas the local linear estimator (2) is not significantly affected by the covariance structure since the data are localized in time (Fan et al., 2007).

It is necessary to derive pointwise confidence intervals for the nonparametric component $\boldsymbol{\alpha}_w(\cdot)$, and to do so we need an estimate of the asymptotic covariance matrix. We use the sandwich estimator

$$\widehat{\mathrm{cov}}\{\hat{\boldsymbol{\alpha}}_w(t_0)\} \approx (\mathbf{I}_p, \mathbf{0}_p)(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\kappa}\boldsymbol{\Lambda})^{-1}\left(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\kappa}\mathcal{Q}\boldsymbol{\kappa}\boldsymbol{\Lambda}\right)(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\kappa}\boldsymbol{\Lambda})^{-1}(\mathbf{I}_p, \mathbf{0}_p)^{\mathrm{T}},$$

where $\mathcal{Q} = \mathrm{diag}(\mathbf{e}_1, \ldots, \mathbf{e}_n)$, with $\mathbf{e}_i = (e_i^2(t_{i1}), \ldots, e_i^2(t_{i\,n_i}))^{\mathrm{T}}$ and $e_i(t) = W_i(t) - \{\mathbf{x}_i^{\mathrm{T}}(t)\hat{\boldsymbol{\alpha}}_w(t) + \mathbf{z}_i^{\mathrm{T}}(t)\hat{\boldsymbol{\beta}}_w\}$.

When the weight matrix $\mathbf{R}$ does not depend on the continuous response $\mathbf{W}$, the estimated covariance matrix for $\hat{\boldsymbol{\beta}}_w$ is obtained using the sandwich formula

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_w) = \mathcal{D}^{-1}\mathcal{V}\mathcal{D}^{-1},$$

where $\mathcal{D} = \mathbf{z}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{S})^{\mathrm{T}}\mathbf{R}(\mathbf{I}_N - \mathbf{S})\mathbf{z}$ and $\mathcal{V} = \mathbf{z}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{S})^{\mathrm{T}}\mathbf{R}\mathcal{Q}\mathbf{R}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{S})\mathbf{z}$.

After we fit a semivarying coefficient model to the continuous response and obtain the residuals from this fit, we move to the second stage. In the second stage we fit a generalized time-varying coefficient model for the conditional model. Cai et al. (2000) introduced generalized varying coefficient models for independent and identically distributed data. We adapt these models to a longitudinal setting.

We start by locally approximating the functions in a neighborhood of a fixed point t_0 via the Taylor expansion:

$$\gamma_r(t) \approx \gamma_r(t_0) + \gamma_r'(t_0)(t - t_0) \equiv a_r^* + b_r^*(t - t_0), \tag{3}$$

for $r = 1, \ldots, p+q+1$. Let $\mathbf{a}^* = (a_1^*, \ldots, a_{p+q+1}^*)^{\mathrm{T}}$ and $\mathbf{b}^* = (b_1^*, \ldots, b_{p+q+1}^*)^{\mathrm{T}}$. For subject i, let $\mathbf{x}_i^*(t) = \left(\mathbf{x}_i^{\mathrm{T}}(t), \mathbf{z}_i^{\mathrm{T}}(t), e_i(t)\right)^{\mathrm{T}}$. We maximize the local likelihood

$$\ell_n(\mathbf{a}^*, \mathbf{b}^*) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ell \left(g^{-1} \left[\sum_{r=0}^{p+q+1} \{a_r^* + b_r^*(t - t_0)\} x_{ir}^*(t) \right], Q_i(t) \right) K_{h_2}(t - t_0), \tag{4}$$

where $g(\cdot)$ is a link function, and h_2 is the bandwidth for the second stage. Since we showed in Section 2.1 that $Q_i(t)$ given $W_i(t)$ follows a probit model, the link function should be probit, in which case (4) becomes

$$\begin{aligned} \ell_n(\mathbf{a}^*, \mathbf{b}^*) = & \frac{1}{N} \sum_{Q_i(t)=1} \log \left(\phi \left[\sum_{r=0}^{p+1} \{a_r^* + b_r^*(t - t_0)\} x_{ir}^*(t) \right] \right) K_{h_2}(t - t_0) \\ & + \frac{1}{N} \sum_{Q_i(t)=0} \log \left(1 - \phi \left[\sum_{r=0}^{p+1} \{a_r^* + b_r^*(t - t_0)\} x_{ir}^*(t) \right] \right) K_{h_2}(t - t_0), \end{aligned} \tag{5}$$

where $\phi(\cdot)$ is the probability density function for the standard normal distribution. We find the solutions to (5) by adapting the iterative local maximum likelihood algorithm described in Cai et al. (2000) to a longitudinal setting, as follows.

Let $a_r^{*(k)}$ and $b_r^{*(k)}$ be the values of a_r^* and b_r^*, respectively, at the kth iteration. Let $\ell_n'(\mathbf{a}^*, \mathbf{b}^*)$ and $\ell_n''(\mathbf{a}^*, \mathbf{b}^*)$ be the gradient and Hessian matrix for the local likelihood (4). Then we update $(\mathbf{a}^*, \mathbf{b}^*)$ according to

$$\begin{pmatrix} \mathbf{a}^{*(k+1)} \\ \mathbf{b}^{*(k+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}^{*(k)} \\ \mathbf{b}^{*(k)} \end{pmatrix} - \left\{ \ell_n''(\mathbf{a}^{*(k)}, \mathbf{b}^{*(k)}) \right\}^{-1} \ell_n'(\mathbf{a}^{*(k)}, \mathbf{b}^{*(k)}).$$

The solution of this iterative regression algorithm satisfies $\ell'(\mathbf{a}^*, \mathbf{b}^*) = 0$, and the estimators are given by $\hat{\mathbf{a}}^* = \hat{\boldsymbol{\gamma}}(t_0) = (\hat{\gamma}_1(t_0), \ldots, \hat{\gamma}_{p+q+1}(t_0))^{\mathrm{T}}$ and $\hat{\mathbf{b}}^* = (\hat{\gamma}_1'(t_0), \ldots, \hat{\gamma}\prime_{p+q+1}(t_0))^{\mathrm{T}}$.

Let $\mathbf{I}$ be the identity matrix with size $p + q + 1$ and $\mathbf{0}$ be a size $p + q + 1$ matrix with each entry equal to zero. Then the asymptotic covariance matrix of the estimator $\hat{\boldsymbol{\gamma}}(t_0)$ can be estimated using the sandwich formula

$$\widehat{\mathrm{cov}}\{\hat{\boldsymbol{\gamma}}(t_0)\} = (\mathbf{I}, \mathbf{0}) \widehat{\boldsymbol{\Gamma}}(t_0)^{-1} \widehat{\boldsymbol{\Delta}}(t_0) \widehat{\boldsymbol{\Gamma}}(t_0)^{-1} (\mathbf{I}, \mathbf{0})^{\mathrm{T}},$$

where

$$\widehat{\mathbf{\Gamma}}(t_0) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\varpi_2\left[\sum_{r=0}^{p+q+1}\left\{\hat{a}_r^*+\hat{b}_r^*(t-t_0)\right\}x_{ir}^*(t), Q_i(t)\right]K_{h_2}(t-t_0)\begin{pmatrix}\mathbf{x}_i^*(t)\\ \mathbf{x}_i^*(t)(t-t_0)\end{pmatrix}^{\otimes 2}$$

$$\widehat{\mathbf{\Delta}}(t_0) = \frac{h}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\varpi_1^2\left[\sum_{r=0}^{p+q+1}\left\{\hat{a}_r^*+\hat{b}_r^*(t-t_0)\right\}x_{ir}^*(t), Q_i(t)\right]K_{h_2}^2(t-t_0)\begin{pmatrix}\mathbf{x}_i^*(t)\\ \mathbf{x}_i^*(t)(t-t_0)\end{pmatrix}^{\otimes 2}$$

with $\varpi_d(\mathcal{A},\mathcal{B}) = (\partial^d/\partial\mathcal{A}^d)\ell\{g^{-1}(\mathcal{A}),\mathcal{B}\}$, and $\mathbf{C}^{\otimes 2}$ denotes $\mathbf{C}\mathbf{C}^{\mathrm{T}}$ for a matrix or vector $\mathbf{C}$.

2 Asymptotics

The asymptotic behavior of our first-stage estimators was presented by Fan et al. (2007). Both estimators are asymptotically normally distributed. The most efficient estimator of $\boldsymbol{\beta}_w$ is obtained when one uses the inverse of the true variance–covariance matrix of the errors as the weight matrix. However, a working independence correlation structure could also be used, in which case the resulting estimate would still be root-n consistent. For the nonparametric component $\boldsymbol{\alpha}_w(t)$, the choice of working correlation structure does not affect the asymptotic bias and variance, which have similar forms to those of the varying coefficient model (Cai et al., 2000).

Kürüm et al. (in press) studied the asymptotic behavior of the estimators in the second stage of our procedure. According to their results, the estimators of the regression coefficients in this stage are asymptotically normally distributed. Note that, in addition to the usual regularity conditions, this result requires the under-smoothing condition $Nh_1 \to 0$. The asymptotic biases of these estimators are also similar to those for varying coefficient models (Cai et al., 2000).

3 Bandwidth Selection

For methods based on kernel smoothing, selecting a suitable bandwidth is an important issue. We recommend using the leave-one-out cross validation method for both stages of our estimation procedure. In a longitudinal study, where intra-subject dependence exists, this approach is more appropriate than leaving out a single observation (Hoover et al., 1998). Specifically, we propose minimizing the following cross validation score:

$$CV(h) = \sum_i \|V_i - \hat{V}_{-i}\|^2,$$

where V_i denotes the observed value of the response V for subject i and $\hat{V}_{-i}$ is the fitted value of this response with subject i excluded. For choosing the first- and second-stage bandwidths, V stands for the continuous and binary responses, respectively. We compute the cross validation score for a range of bandwidths and select the bandwidth that minimizes the score.

We suggest employing a bimodal kernel (De Brabanter et al., 2011) to obtain more accurate estimates in the presence of intra-subject dependence. A bimodal kernel prevents undersmoothing by removing serial dependence through down weighting observations that are very close to t_0. We use a member of the so called ε-optimal class of bimodal kernels suggested by De Brabanter et al. (2011). Specifically, we use

$$K_\varepsilon(u) = \frac{4}{4 - 3\varepsilon - \varepsilon^2} \begin{cases} \frac{3}{4}(1-u^2)\mathbf{1}\{|u| \leq 1\} & \text{if } |u| \geq \varepsilon \\ \frac{3}{4}\frac{1-\varepsilon^2}{\varepsilon}|u| & \text{if } |u| < \varepsilon \end{cases}$$

with $\varepsilon = 0.1$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

References

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):888–902.

Cleveland, W., Grosse, E., and Shyu, W. (1992). *Local regression models.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

De Brabanter, K., De Brabanter, J., Suykens, J., and Moor, B. D. (2011). Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research*, 12:1955–1976.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of American Statistical Association*, 102(478):632–641.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B*, 55(4):757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Kürüm, E., Li, R., Shiffman, S., and Yao, W. (in press). Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. *Statistica Sinica*.