A unifying perspective on smoothing, mixed models and correlated data

Thomas Kneib

Faculty of Mathematics and Economics, University of Ulm
Department of Statistics, Ludwig-Maximilians-University Munich

joint work with
Stefan Lang (University of Innsbruck)

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

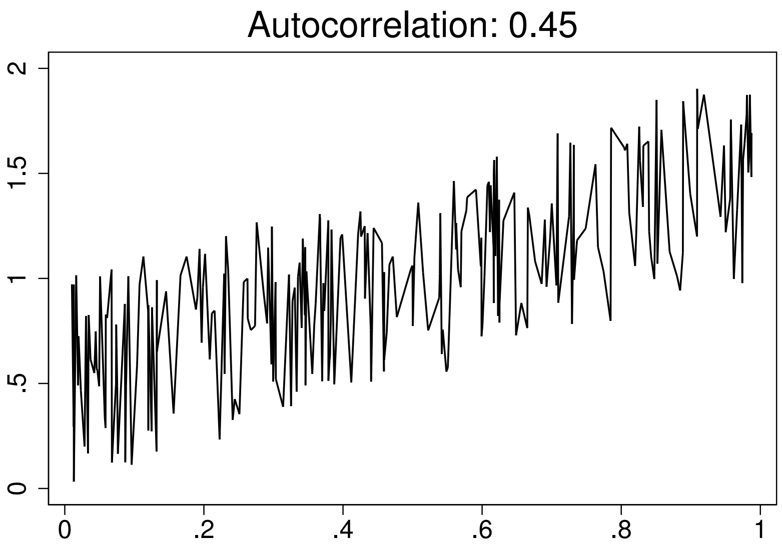
19.7.2007

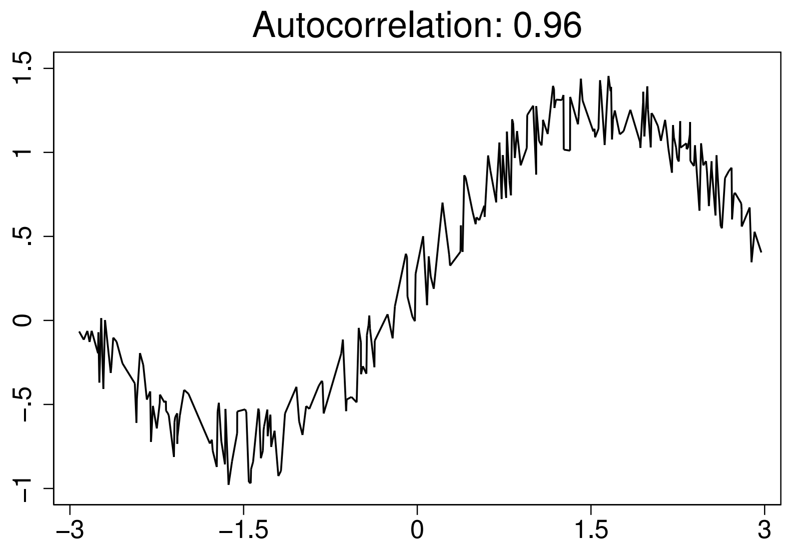
What is Correlation?

- Development economics is often faced with data evolving in both time and space.
- Statistical analyses have to take the special structure into account, i.e.
  - account for spatio-temporal correlations,
  - account for space- and time-varying effects,
  - model unobserved heterogeneity due to spatial and temporal variation.
- Are these really different tasks or merely different phrases for the same goal?

- What is (positive) correlation?

  $\Rightarrow$ Observations which are positively correlated behave "similar".





- Correlation is commonly assumed to be a stochastic phenomenon.

- The above data have been generated from deterministic models:

$$y_t = t + \varepsilon_t \qquad\qquad y_t = \sin(t) + \varepsilon_t$$

- Temporal correlation is often (at least partly) attributable to a trend function.
- The trend itself is typically introduced by unobserved, temporally / spatially varying covariates.
- Usually the response is not influenced by time or space directly (no causal relationship).

Mixed Models I: Classical Perspective

- Longitudinal data: Repeated measurements

$$y_{it}, \quad i = 1, \dots, n, \quad t = 1, \dots, T$$

on a fixed set of subjects $i = 1, \dots, n$ at time points $t = 1, \dots, T$.

- Classical model for such data: Mixed effects / random effects models.

- Simplest example: Random intercepts

$$y_{it} = x_{it}'\beta + b_i + \varepsilon_{it}$$

where

$$b_i \quad \text{i.i.d.} \quad N(0, \tau^2),$$
$$\varepsilon_{it} \quad \text{i.i.d.} \quad N(0, \sigma^2).$$

- Two sources of random variation: Variation on the subject level (b_i) and variation on the measurement level (ε_{it}).

- Rationale: The observations i are a random sample from the population of individuals.

- The random effects distribution b_i i.i.d. $N(0, \tau^2)$ describes the distribution of individual-specific effects b_i in this population.

- Corresponding density:

$$p(b) \propto \exp\left(-\frac{1}{2\tau^2} b'b\right)$$

  where $b = (b_1, \dots, b_n)'$.

- Estimation in mixed models is based on the joint likelihood

$$\begin{aligned} p(y,b) &= p(y|b)p(b) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta - Zb)'(y - X\beta - Zb)\right)\exp\left(-\frac{1}{2\tau^2}b'b\right) \to \max_{\beta,b}. \end{aligned}$$

- Equivalently, we can consider the joint least-squares criterion

$$(y - X\beta - Zb)'(y - X\beta - Zb) + \frac{\sigma^2}{\tau^2}b'b \to \min_{\beta,b}.$$

Mixed Models II: Marginal Perspective

- Hierarchical formulation of mixed models:

$$\begin{aligned} y_{it}|b_i &\sim N(x_{it}'\beta + b_i, \sigma^2) \\ b_i &\sim N(0, \tau^2). \end{aligned}$$

- What happens, if we marginalize with respect to the b_i?

  $\Rightarrow$ Correlation between observations on one individual are induced due to the shared random effects b_i.

- To be more specific: An equicorrelation model is obtained

$$\mathrm{Corr}(y_{it_1}, y_{it_2}) = \frac{\mathrm{Var}(b_i)}{\mathrm{Var}(b_i) + \mathrm{Var}(\varepsilon_{it})} = \frac{\tau^2}{\tau^2 + \sigma^2} = \rho,$$

- Marginal model in matrix notation:

$$y_i \sim N(X_i\beta, \Sigma_i),$$

where

$$\Sigma_i = (\sigma^2 + \tau^2) \begin{pmatrix} 1 & \rho & \dots & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \rho \\ \rho & \dots & \dots & \rho & 1 \end{pmatrix}.$$

Mixed Models III: Penalised Likelihood Perspective

- Start with the model equation

$$y_{it} = x_{it}'\beta + b_i + \varepsilon_{it}$$

without a distributional assumption for b_i.

- The b_i are individual-specific regression coefficients that shall capture effects of unobserved, individual-specific covariates.

- The number of these effects is large

$\Rightarrow$ Add a ridge penalty to stabilise estimation.

- Instead of the least squares criterion

$$(y - X\beta - Zb)'(y - X\beta - Zb) \to \min_{\beta,b}$$

we minimise the penalised least squares criterion

$$(y - X\beta - Zb)'(y - X\beta - Zb) + \lambda b'b \to \min_{\beta,b}$$

- The penalty shrinks parameters b_i to zero, in particular if the database for individual i is small.

- The penalised least squares criterion is equivalent to the joint likelihood of the mixed model with

$$\lambda = \frac{\sigma^2}{\tau^2},$$

i.e. the error to signal ratio determines the strength of the penalisation.

Mixed Models IV: Bayesian Perspective

- Bayesian view: The random effects distribution can be considered as a prior distribution that expresses our knowledge about the individual-specific effects.

- $b_i \sim N(0, \tau^2)$ a priori implies that
  - we expect the effects to be "not too far" from zero,
  - we expect the family of effects in the population to be Gaussian.

  $\Rightarrow$ Qualitatively similar to the random effects view.

- No formal differentiation between fixed and random effects: Both are random quantities but with different a priori knowledge.

$$p(\beta) \propto \text{const} \qquad p(b) \propto \exp\left(-\frac{1}{2\tau^2} b'b\right)$$

- Estimation is based on the posterior

$$p(\beta, b|y) = \frac{p(y|\beta, b)p(\beta)p(b)}{p(y)} \propto p(y|\beta, b)p(b).$$

- The posterior mode coincides with the penalised least squares estimate.

Mixed Models V: Summary

- Four views on the model
$$y_{it} = x_{it}'\beta + b_i + \varepsilon_{it}$$
for longitudinal data:
  - Mixed model perspective: b_i is a random effect from the population distribution.
  - Marginal perspective: the b_i induce equicorrelation.
  - Penalised likelihood perspective: the b_i are individual-specific regression coefficients.
  - Bayesian perspective: the random effects distribution expresses a priori knowledge.
- Both the mixed model and the Bayesian perspective combine features of the two further perspectives.
- Different rationales but the same goal: Describe / analyse why observations of one individual behave more similar than randomly selected measurements.

- What do we gain by the different perspectives:
  - Different estimation schemes have been developed by the different statistical communities.
  - Additional insight in more complicated types of models, e.g. concerning identifiability problems when modelling both trend functions and correlation.

Mixed Models VI: Extensions

- Similar considerations can be made for extended models such as
  - Models with random slopes:

$$y_{it} = x_{it}'\beta + z_{it}'b_i + \varepsilon_{it}.$$

  - Nested multi-level models

$$y_{ijt} = x_{ijt}'\beta + b_i + b_{ij} + \varepsilon_{ijt}.$$

  - Non-Nested multi-level models

$$y_{ijt} = x_{ijt}'\beta + b_i + b_j + \varepsilon_{ijt}.$$

Smoothing and Mixed Models

- Consider trend estimation in the simple model

$$y_t = f_{\text{trend}}(t) + \varepsilon_t, \qquad \varepsilon_t \text{ i.i.d. } N(0, \sigma^2).$$

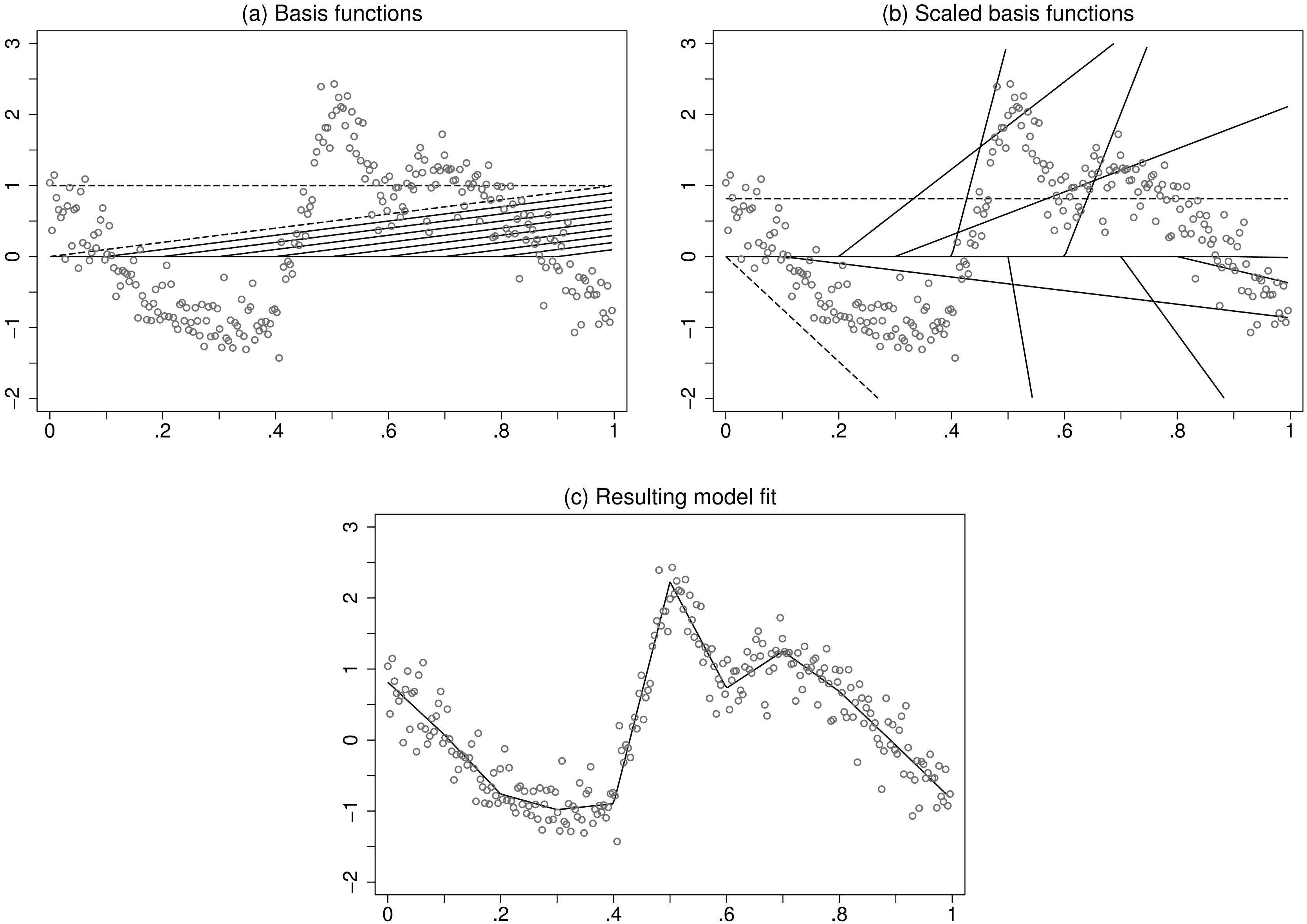
- Model the trend function as a polynomial spline (in truncated line representation):

$$f_{\text{trend}}(t) = \beta_0 + \beta_1 t + b_1(t - \kappa_1)_+ + \ldots + b_d(t - \kappa_d)_+.$$

$\Rightarrow$ Piecewise linear function estimate with changing slopes at the knots κ_j.

- In matrix notation

$$y = X\beta + Zb + \varepsilon.$$


(a) Basis functions
(b) Scaled basis functions
(c) Resulting model fit
3
2
1
0
−1
−2
0
.2
.4
.6
.8
1

- To avoid overfitting, introduce a penalty term for the truncated polynomials:

$$\lambda \sum_{j=1}^{d} b_j^2 = \lambda b' b.$$

$\Rightarrow$ Variability of the function estimate is controlled by the smoothing parameter λ.

- λ large $\Rightarrow$ $\hat{f}(x)$ approaches a linear function.

- λ small $\Rightarrow$ $\hat{f}(x)$ becomes a very wiggly estimate.

- Estimate the parameters of the trend function by minimising the penalised least squares criterion

$$(y - X\beta - Zb)'(y - X\beta - Zb) + \lambda b'b \to \min_{\beta,b}$$

with smoothing parameter λ.

- This is the same objective function as for a mixed model

$$y = X\beta + Zb + \varepsilon$$

with distributional assumptions

$$\begin{bmatrix} \varepsilon \\ b \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \tau^2 I \end{bmatrix} \right)$$

where $\lambda = \sigma^2/\tau^2$.

$\Rightarrow$ The smoothing approach for trend estimation can be considered a mixed model with very specific structure.

- Consequences:
  - Mixed model methodology can be used to estimate the smoothing parameter λ (the ratio of error variance and random effects variance).
  - Conditionally on b we are modelling a trend function but marginally the model implies correlation of the response.

    $\Rightarrow$ Simultaneous modelling of trend functions and correlated errors may cause identifiability problems.
  - All four perspectives can be applied to the model, yielding for example a Bayesian interpretation.

Autoregressive Processes as Smoothers

- Consider the model

$$y_{it} = x_{it}'\beta + b_t + \varepsilon_{it}$$

where ε_{it} i.i.d. $N(0, \sigma^2)$ and b_t follows an autoregressive process of order 1 (AR(1))

$$b_t = \alpha b_{t-1} + u_t, \quad u_t \sim N(0, \tau^2).$$

- Note: b_t is now a temporally correlated effect, not an individual-specific effect.

- Correlation function of the autoregressive process (with parameter α):

$$\rho(b_t, b_s) = \alpha^{|t-s|}.$$

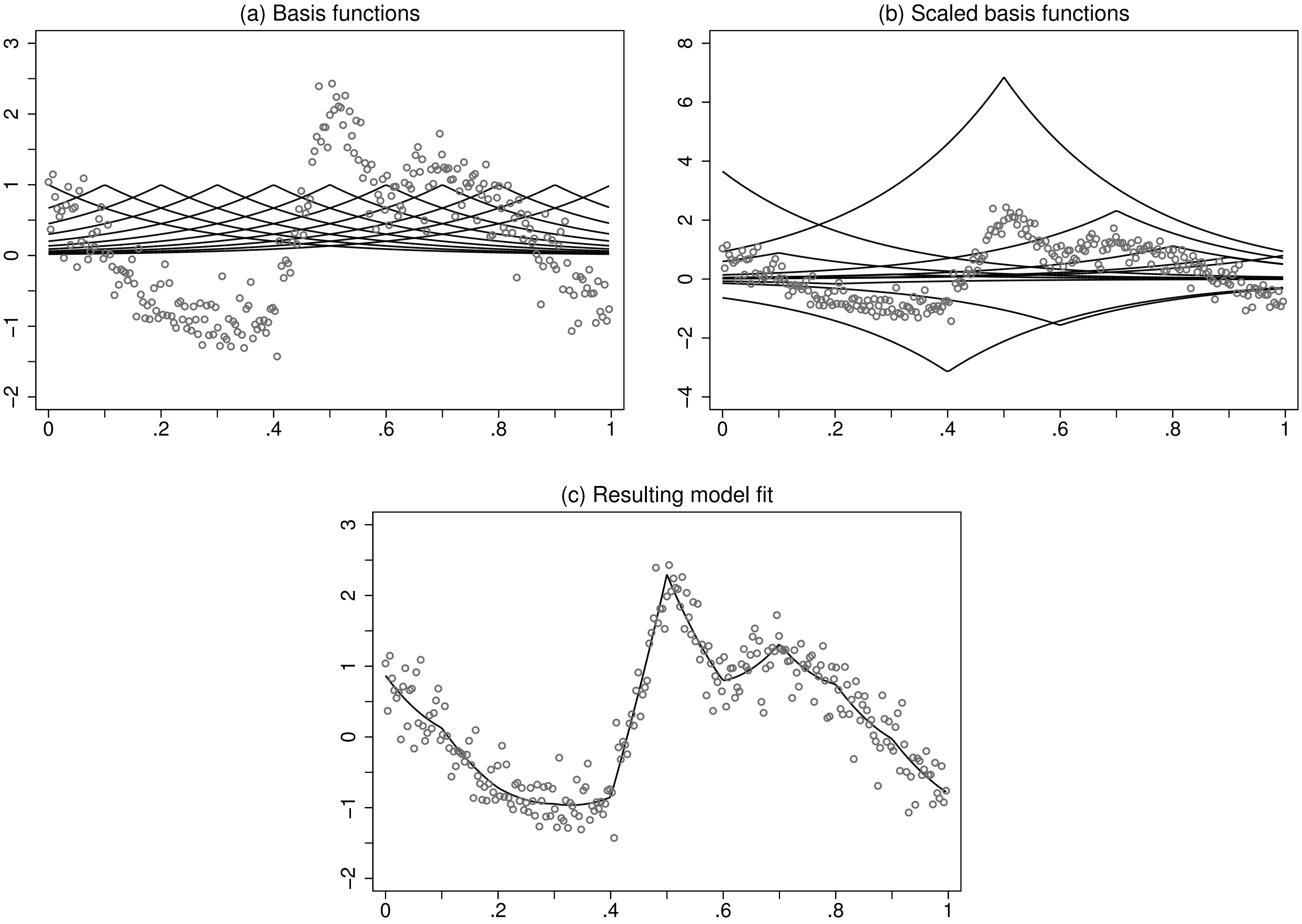
- This is a correlation function in discrete time. The continuous time analogue is the exponential correlation function

$$\rho(b_t, b_s) = \exp\left(-\frac{|t-s|}{\phi}\right), \qquad \alpha = \exp\left(-\frac{1}{\phi}\right)$$

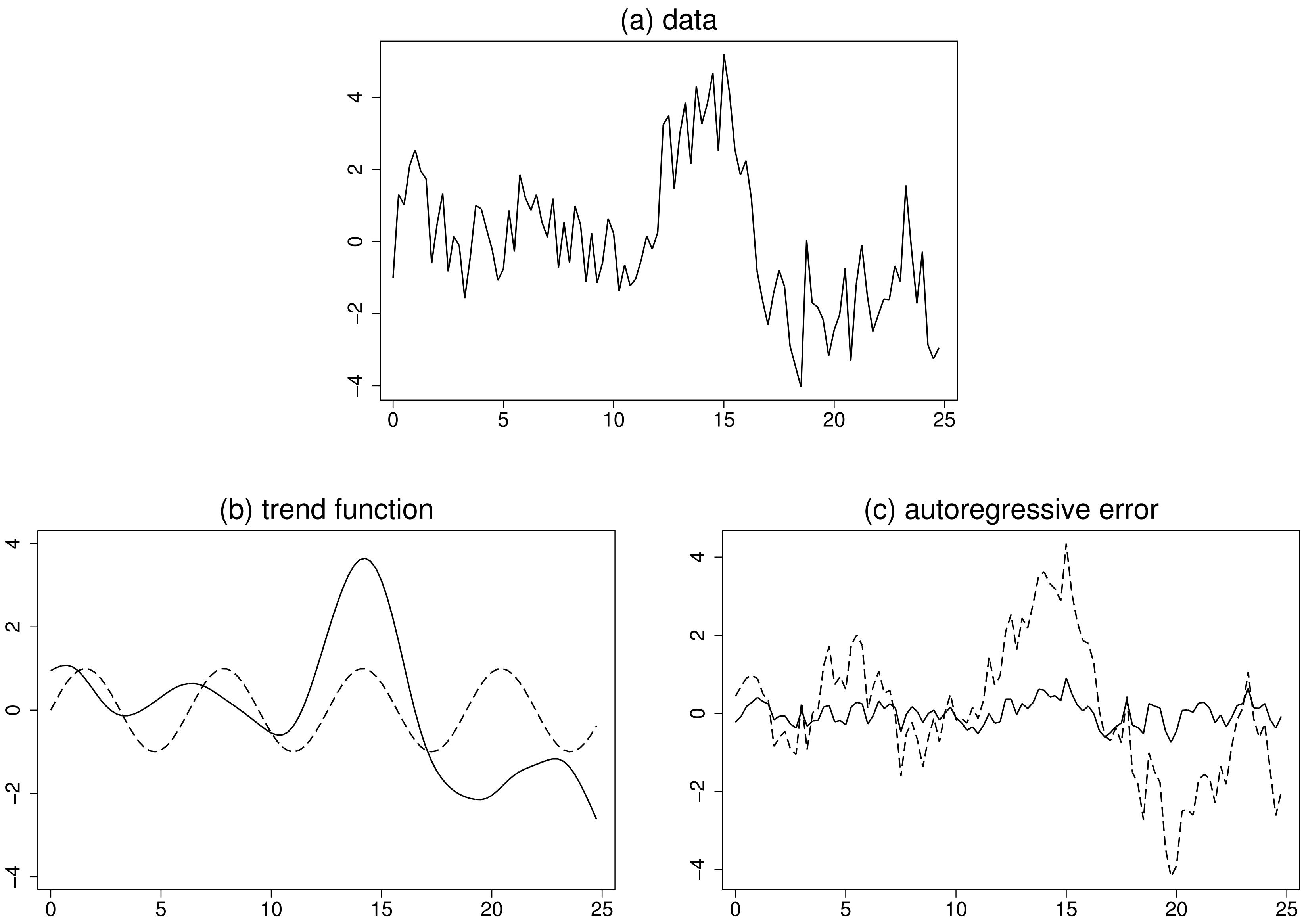
- It can be shown that the temporally correlated effect can be rewritten as

$$b_t = f(t) = \sum_{s=1}^{T} \rho(b_t, b_s)\gamma_t.$$

$\Rightarrow$ The AR(1) assumption is equivalent to a basis function approach.


(a) Basis functions
(b) Scaled basis functions
(c) Resulting model fit
0
.2
.4
.6
.8
1

- Consequences:
  - The AR(1) correlation function can be interpreted as a (radial) basis function.
  - A similar relation holds for stochastic processes with different types of correlation functions.
  - The autoregressive process assumption turns into a penalty for the parameter vector γ_t.
  - The result can be immediately extended to spatial models with spatially autoregressive errors and spatial trend functions.
  - The larger the autoregressive parameter, the smoother the basis function.
  - Identifiability problems when including both a highly correlated autoregressive error and a flexible trend function.


(a) data
(b) trend function
(c) autoregressive error

A Unifying Framework

- Structured additive regression:
  - Combines nonparametric regression, spatial regression, random effects, etc.
  - General model equation:

$$y = f_1(z_1) + \ldots + f_r(z_r) + x'\beta.$$

  - Examples:

$f(z) = f(x)$	$z = x$	smooth function of a continuous covariate x,
$f(z) = f_{\text{spat}}(s)$	$z = s$	spatial effect,
$f(z) = f(x_1, x_2)$	$z = (x_1, x_2)$	interaction surface,
$f(z) = b_g$	$z = g$	i.i.d. frailty b_g, g is a grouping index.

  - Can be extended to non-Gaussian responses.

- Generic representation of the different effect types:
  - Vectors of function evaluations:

$$f_j = Z_j \gamma_j$$

  - Prior distribution / random effects distribution / penalty term:

$$p(\gamma) \propto \exp\left(-\frac{1}{2\tau^2}\gamma' K_j \gamma\right), \qquad \text{Pen}(\gamma) = \lambda \gamma' K_j \gamma.$$

- Four different perspectives:
  - Penalised likelihood setting:

$$\left(y - X\beta - \sum_{j=1}^{r} Z_j\gamma_j\right)'\left(y - X\beta - \sum_{j=1}^{r} Z_j\gamma_j\right) + \sum_{j=1}^{r} \lambda_j \gamma_j' K_j \gamma_j \rightarrow \min_{\beta,\gamma_1,\dots,\gamma_r}$$

  - Mixed model perspective: The γ_j are correlated random effects. Estimation is based on the joint likelihood

$$p(y|\gamma_1,\dots,\gamma_r)p(\gamma_1,\dots,\gamma_r) \rightarrow \max_{\beta,\gamma_1,\dots,\gamma_r}$$

  - Bayesian view: The mixed model distribution defines a prior for γ_j.
  - Marginal view: After integrating out the random effects γ_j, we obtain a marginal model

$$y \sim N(X\beta, V),$$

where V is a covariance matrix with correlations induced by the random effects.

Conclusions

- Four different perspectives on semiparametric regression.
- Though looking different at first sight, there is a close connection between all them.
- In particular, semiparametric smoothing and modelling of correlations are related tasks.
- Identifiability problems can be encountered when flexibly modelling correlations and temporal / spatial trend functions.
- The different perspectives allow to derive different estimation techniques.