Recent Advances in Linear Models and Related Areas

Shalabh • Christian Heumann

Recent Advances in Linear Models and Related Areas

Essays in Honour of Helge Toutenburg

Springer

Dr. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Kanpur - 208016 India
shalab@iitk.ac.in

Dr. Christian Heumann
Institute of Statistics
Ludwig-Maximilians-University Munich
Akademiestr. 1
80799 Munich
chris@stat.uni-muenchen.de

ISBN 978-3-7908-2063-8 e-ISBN 978-3-7908-2064-5

Library of Congress Control Number: 2008929501

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

This collection contains invited papers by distinguished statisticians to honour and acknowledge the contributions of Professor Dr. Dr. Helge Toutenburg to Statistics on the occasion of his sixty-fifth birthday. These papers present the most recent developments in the area of the linear model and its related topics.

Helge Toutenburg is an established statistician and currently a Professor in the Department of Statistics at the University of Munich (Germany) and Guest Professor at the University of Basel (Switzerland). He studied Mathematics in his early years at Berlin and specialized in Statistics. Later he completed his dissertation (Dr. rer. nat.) in 1969 on optimal prediction procedures at the University of Berlin and completed the post-doctoral thesis in 1989 at the University of Dortmund on the topic of mean squared error superiority. He taught at the Universities of Berlin, Dortmund and Regensburg before joining the University of Munich in 1991.

He has various areas of interest in which he has authored and co-authored over 130 research articles and 17 books. He has made pioneering contributions in several areas of statistics, including linear inference, linear models, regression analysis, quality engineering, Taguchi methods, analysis of variance, design of experiments, and statistics in medicine and dentistry. His most influential contributions are in the area of optimal prediction in linear models, mean squared error superiority of biased estimators, weighted mixed estimation in missing data analysis, repeated measures designs and the unification of various parameterizations of the carry-over effect in cross-over designs. His books *Prediction and Improved Estimation in Linear*

Models (Wiley) and *Prior Information in Linear Models* (Wiley) laid the foundations for further work in the field of utilization of prior information as well as in the field of prediction. Other pioneering works include *Linear Models and Generalizations: Least Squares and Alternatives* (Springer) and *Statistical Analysis of Designed Experiments* (Springer). His books in German on descriptive and inductive statistics, quality engineering, design of experiments and linear models are among the popular textbooks in several universities in Germany. He has also translated the celebrated books of Professor C.R. Rao into German. His book on statistics in dentistry is the first book in German in this area.

Helge Toutenburg maintains fruitful research collaboration with researchers in different countries like the USA, India, Korea, etc. He has hosted DAAD and Humboldt fellows. He is not only a well known researcher but also an excellent teacher. He has advised Ph.D. students from germany and abroad. His efficient working style has always been appreciated by those who had a chance to collaborate with him. He has been actively associated with the International Statistical Institute, Deutscher Hochschulverband, Deutsche Statistische Gesellschaft, Biometrical Society and Bernoulli Society for Mathematical Statistics and Probability.

Besides having a great interest in statistics, Helge Toutenburg has a great sense of humor, too. He has written several books on humor in German to the pleasure of his friends and colleagues.

This collection of invited papers brings together the recent developments in the field of linear models and its related sub-fields as well as papers from Helge Toutenburg‘s other areas of interest.

As the editors of this book, we would like to express our heartful thanks to the authors whose contributions and commitment made this book possible. We would like to thank Michael Schomaker for his immense help in the editorial process and Valentin Wimmer for his help in typing. We are also thankful to Dr. Müller of Springer for his cooperation in the publication of this book.

Munich, *Shalabh*
June, 2008 *Christian Heumann*

Contents

List of Contributors

A.K.Md. Ehsanes Saleh
Carleton University/School of Mathematics and Statistics
Ottawa, Canada
esaleh@math.carleton.ca

Alan T.K. Wan City University of Hong Kong/Department of Management Sciences - Kowloon, Hong Kong
msawan@cityu.edu.hk

Aman Ullah University of California/Department of Economics University Avenue 900 - Riverside, CA 92521, U.S.A.
aman.ullah@ucr.edu

Angela Döring GSF-National Research Center for Environment and Health, Ingolästter Landstrasse 1, 85764 Neuherberg, Germany
doering@gsf.de

Anoop Chaturvedi University of Allahabad/Department of Statistics - Allahabad, India
anoopchaturv@gmail.com

Bettina Grün Technische Universität Wien/Institut für Statistik und Wahrscheinlichkeitstheorie Wiedner Hauptstrasse 8-10 - 1040 Vienna, Austria
Bettina.Gruen@ci.tuwien.ac.at

Burkhard Schaffrin Ohio State University/School of Earth Sciences - Columbus Ohio, U.S.A.
schaffrin.1@osu.edu

C. Radhakrishna Rao Penn State University/Department of Statistics PA 16802 - University Park, U.S.A.
crr1@psu.edu

Chandra Mani Paudel Tribhuvan University/Department of Statistics - Pokhra, Nepal
cpaudel@hotmail.com

Changli He Dalarna University - Dalarna, Sweden
chh@du.se

Christian Heumann University of Munich/Department of Statistics Akademiestrasse 1 - 80799 Munich, Germany
christian.heumann@stat.uni-muenchen.de

Christoph Hanck University of Dortmund/Department of Statistics Vogelpothsweg 78 - 44221 Dortmund, Germany
christoph.hanck@uni-dortmund.de

David Rummel emnos GmbH Theresienhöhe 12 - 80339 Munich, Germany d_rummel@web.de

Dietrich Trenkler University of Osnabrück/Rolandstrasse 8 - 49069 Osnabrück, Germany
d.trenkler@nts6.oec.uni-osnabrueck.de

Friedrich Leisch University of Munich/Department of Statistics Ludwigstrasse 33 - 80539 Munich, Germany
Friedrich.Leisch@stat.uni-muenchen.de

Guoqi Qian University of Melbourne/Department of Mathematics and Statistics VIC 3010 - Melbourne, Australia
g.qian@ms.unimelb.edu.au

Gerhard Tutz University of Munich/Department of Statistics Akademiestrasse 1 - 80799 Munich, Germany
tutz@stat.uni-muenchen.de

Götz Trenkler University of Dortmund/Department of Statistics Vogelpothsweg 87 - 44221 Dortmund, Germany
trenkler@statistik.uni-dortmund.de

Guohua Zou University of Rochester/Department of Biostatistics and Computational Biology Rochester, NY 14642, U.S.A.
Guohua_Zou@URMC.Rochester.edu

Hans Malmsten Länsförsäkringar - Stockholm, Sweden Hans.Malmsten@lansforsakringar.se

Hans Schneeweiss University of Munich/Department of Statistics Ludwigstrasse 33 - 80539 Munich, Germany
schneew@stat.uni-muenchen.de

Helmut Küchenhoff University of Munich/Department of Statistics Akademiestrasse 1 - 80799 Munich, Germany
helmut.kuechenhoff@stat.uni-muenchen.de

Hilmar Drygas University of Kassel/Department of Mathematics - 34109 Kassel, Germany
drygas@mathematik.uni.kassel.de

Huaizhen Qin Michigan Technological University/Department of Mathematical Sciences 1400 Townsend Drive - Houghton, Michigan 49931-1295, U.S.A.
hqin@mtu.edu

Hyang Mi Kim University of Calgary/Department of Mathematics and Statistics - Calgary, Canada
hmkim@ucalgary.ca

Hyuk Joo Kim Wonkwang University/Division of Mathematics - Jeonbuk 570-749, Korea
hjkim@wonkwang.ac.kr

Iris Pigeot Bremen Institute for Prevention Research and Social Medicine Linzer Strasse 10 - 28359 Bremen, Germany
pigeot@bips.uni-bremen.de

Jae-Il Cho Dongbu Electronics/Management Innovation
Part Street, Gyeonggi 420-712, Korea
jaeil.cho@dsemi.com

Jan Ulbricht University of Munich/Department of Statistics Akademiestrasse 1 - 80799 Munich, Germany
ulbricht@stat.uni-muenchen.de

Jörg Drechsler Institute for Employment Research of the Federal Employment Agency/Department for Statistical Methods Regensburger Strasse 104 - 90478 Nürnberg, Germany
joerg.drechsler@iab.de

Ludwig Fahrmeir University of Munich/Department of Statistics Ludwigstrasse 33 - 80539 Munich, Germany
ludwig.fahrmeir@stat.uni-muenchen.de

Michael Akritas Penn State University/Department of Statistics 325, Thomas Building, University Park - PA 16802, U.S.A.
mga@stat.psu.edu

Narinder Kumar Panjab University/Department of Statistics - Chandigarh 160014, India
nkumar@pu.ac.in

Nina Wawro Bremen Institute for Prevention Research and Social Medicine Linzer Strasse 10 - 28359 Bremen, Germany
wawro@bips.uni-bremen.de

Panu Somervuo University of Helsinki/Institute of Biotechnology P.O.Box 56 - Helsinki 00014, Finland
pjsomerv@mappi.helsinki.fi

Petri Auvinen University of Helsinki/Institute of Biotechnology P.O.Box 56 - Helsinki 00014, Finland
petri.auvinen@helsinki.fi

Qing Shao Novartis Pharmaceuticals
Corporation/Biostatistics and Statistical Reporting One Health Plaza, Bldg. 435 - East Hanover, NJ 07936, U.S.A.
qing.shao@novartis.com

Ram Chandra University of Agriculture and Technology - Kanpur, India ramchandra20@rediffmail.com

Rashi Gupta University of Helsinki/Department of Mathematics and Statistics P.O. Box 68 - Helsinki 00014, Finland
rashi.gupta@helsinki.fi

Sangita Kulathinal University of Helsinki/Department of Mathematics and Statistics P.O. Box 68 - Helsinki 00014, Finland
sangita.kulathinal@inseed.org

Shalabh Indian Institute of Technology/Department of Mathematics and Statistics - Kanpur 208016, India
shalab@iitk.ac.in

Shu-Min Liao Penn State University/Department of Statistics 325, Thomas Building, University Park - PA 16802, U.S.A.
sx1340@psu.edu

Suchita Kesarwani University of Allahabad/Department of Statistics - Allahabad, India
suchitakesarwani@yahoo.co.in

Sung Hyun Park Seoul National University/
Department of Statistics - Seoul 151-747, Korea
parksh@plaza.snu.ac.kr

Susanne Rässler Otto-Friedrich-University
Bamberg/Department of Statistics and Econometrics
Feldkirchenstrasse 21 - 96045 Bamberg, Germany
susanne.raessler@sowi.uni-bamberg.de

Thomas Augustin University of Munich/Department of Statistics Ludwigstrasse 33 - 80539 Munich, Germany
thomas@stat.uni-muenchen.de

Thomas Kneib University of Munich/Department of Statistics Ludwigstrasse 33 - 80539 Munich, Germany
thomas.kneib@stat.uni-muenchen.de

Thomas Nittner UBS AG/Credit & Country Risk Controlling
Pelikanstrasse 6/8 - 8098 Zurich, Switzerland
thomas.nittner@ubs.com

Timo Teräsvirta University of Aarhus/CREATES, School of Economics and Management - Aarhus, Denmark
tterasvirta@econ.au.dk

Walter Krämer University of Dortmund/Department of Statistics Vogelpothsweg 78 - 44221 Dortmund, Germany
walterk@statistik.uni-dortmund.de

Weiqiang Qian University of California/Department of Economics University Avenue 900 - Riverside, CA 92521, U.S.A.
weiqiang.qian@ucr.edu

Yuehua Wu York University/Department of Mathematics and Statistics 4700 Keele Street - Toronto, M3J 1P3, Canada
wuyh@yorku.ca

On the Identification of Trend and Correlation in Temporal and Spatial Regression

Ludwig Fahrmeir[1] and Thomas Kneib[2]

[1] Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany ludwig.fahrmeir@stat.uni-muenchen.de
[2] Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany thomas.kneib@stat.uni-muenchen.de

1 Introduction

In longitudinal or spatial regression problems, estimation of temporal or spatial trends is often of primary interest, while correlation itself is of secondary interest or is regarded as a nuisance component. In other situations, the stochastic process inducing the correlation may be of interest in itself. In this paper, we investigate for some simple time series and spatial regression models, how well trend and correlation can be separated if both are modeled in a flexible manner.

From a classical point of view, trends are considered as deterministic unknown functions to be estimated from the data, whereas correlation is thought to be generated from an unobservable, latent temporal or spatial process. If the focus of statistical inference is on recovering trends, then the latent error process is often only used to give some guidance in choosing reasonable correlation functions to enhance quality of trend estimation. Even more, the derived correlation structure may only be considered as working correlation such as in marginal models for longitudinal data, see e.g. Toutenburg (2003, Ch. 10) and the references therein. To make the discussion concrete, let us consider a simple nonparametric regression problem, where observations $y(t_i)$ on a process $\{y(t), t \geq 0\}$ are available at time points $t_1 < \ldots < t_n$, say. The observable process is related to an unknown trend function $f(t)$ through the additive relation

$$y(t_i) = f(t_i) + \varepsilon(t_i), \quad i = 1, \ldots, n \ , \tag{1}$$

where $\varepsilon(t), t \geq 0$ is an unobservable Gaussian error process with marginal distributions $\varepsilon(t) \sim N(0, \sigma^2)$. Defining the vectors $y =$

$(y(t_1), \ldots, y(t_n))'$, $f = (f(t_1), \ldots, f(t_n))'$ and $\varepsilon = (\varepsilon(t_1), \ldots, \varepsilon(t_n))'$, we obtain the model in matrix notation as

$$y = f + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 R), \tag{2}$$

where the correlation matrix R has elements $r_{ij} = \rho(\varepsilon(t_i), \varepsilon(t_j))$ with some suitable correlation function ρ. In a purely parametric approach, the trend function could be approximated as a linear combination

$$f(t) = \sum_{j=1}^{p} \beta_j B_j(t) \tag{3}$$

of a few basis functions. To achieve optimality, the unknown coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ would then be estimated by minimizing a weighted least squares criterion based on the 'true' correlation matrix R or a consistent estimate $\hat{R}$.

Simple parametric forms like (3) are often too restrictive, at least prior to exploratory data analysis, for modelling trend functions. The most popular nonparametric alternatives are basis function approaches in combination with penalization, such as smoothing splines or penalized splines, and kernel-based local regression techniques. In case of i.i.d errors ε_t, where $R = I$, there is a close connection between both concepts, see e.g. Fahrmeir and Tutz (2001, Ch. 5), and empirical experience shows that they often lead to rather comparable estimates from a practical point of view.

It might intuitively be expected that this similarity in practical performance transfers to the case of correlated error processes as long as a good estimate of R is available. Surprisingly, this is not the case. Kohn, Schimek and Smith (2000) point out some emerging yet different consequences if correlation is neglected in estimation procedures, and they suggest some remedies. Lin and Carroll (2000) show that common kernel-based methods work best when correlation is neglected, i.e if $R = I$ is used as a working correlation matrix. Welsh, Lin and Carroll (2002) provide additional support for this result, but they also confirm that efficient spline estimates are obtained when using the true correlation structure. As a reaction to these somewhat surprising results, Wang (2003) and Linton, Mammen, Lin and Carroll (2004) constructed modified kernel-based estimates which improve upon the usual kernel estimates. Krivobokova and Kauermann (2007) investigate penalized spline estimation for time series data within a mixed model framework and provide some evidence that relatively robust nonparametric estimates are obtained when smoothing parameters are cho-

sen as restricted maximum likelihood estimates even if the correlation structure is misspecified.

In this contribution we shed some further light on this puzzle from a Bayesian perspective. We focus on approaches with Bayesian smoothing priors for modeling trend functions, such as random walk models or extensions to Bayesian penalized (P-)splines. If the correlation-generating error process has similar stochastic structure as the smoothing prior it seems quite plausible that identifiability problems can arise. In particular, it can become difficult to separate trend from correlation. We first exemplify this using a simple time series setting in Section 2. In Section 3 we move on to the corresponding spatial situation, which arises in geostatistics. Section 4 briefly points out extensions to the general class of structured additive regression (STAR) models.

2 Trend and Correlation in Time Series Regression

Let us first revisit the classical smoothing problem already treated by Whittaker (1923), which is closely related to the nonparametric regression problem (2). Time series observations $y(t)$ on an equidistant grid of time points $t = 1, \ldots, n$ are assumed to be the sum

$$y(t) = f(t) + \varepsilon(t), \quad t = 1, \ldots, n \tag{4}$$

of a smooth trend function f and an irregular noise component ε with i.i.d. errors $\varepsilon \sim N(0, \sigma^2)$. Whittaker suggested to estimate f by minimizing the penalized least squares (PLS) criterion

$$\mathrm{PLS}(f) = \sum_{t=1}^{n} (y(t) - f(t))^2 + \lambda \sum_{t=d+1}^{n} (\Delta^d f(t))^2 \tag{5}$$

where λ is a given smoothing parameter, and the sum of (squared) first ($d = 1$) or second ($d = 2$) order differences

$$\Delta^1 f(t) = f(t) - f(t-1), \qquad \Delta^2 f(t) = f(t) - 2f(t-1) + 2f(t),$$

penalizes deviations from a horizontal or a straight line, respectively. In matrix notation, the observation model becomes $y = f + \varepsilon$ as in (2), and the penalized least squares criterion (5) can be expressed as

$$\mathrm{PLS}(f) = (y - f)'(y - f) + \lambda f' K_d f, \tag{6}$$

with penalty matrix $K_d, d = 1, 2$, given by

$$K_d = D_d' D_d \tag{7}$$

where D_1 and D_2 are first and second order difference matrices, respectively. It can be easily shown that

$$\hat{f} = (I + \lambda K_d)^{-1} y \tag{8}$$

minimizes PLS(f). The (frequentist) covariance matrix of the PLS-estimate is given by

$$\mathrm{Cov}(\hat{f}) = \sigma^2 (I + \lambda K_d)^{-2}\,. \tag{9}$$

The *Bayesian version* of the smoothing problem of Whittaker can be formulated as a hierarchical model consisting of two stages. Assuming i.i.d. Gaussian errors $\varepsilon(t) \sim N(0, \sigma^2)$, the first stage is the observation model

$$y|f \sim N(f, \sigma^2 I).$$

The second stage specifies a smoothness prior for the unknown function, more exactly for the vector $f = (f(1), \ldots, f(n))'$ of function values. The stochastic analogue of first or second order difference penalties are random walk priors of first ($RW(1)$) or second ($RW(2)$) order

$$f(t) = f(t-1) + u(t)$$

or

$$f(t) = 2f(t-1) - f(t-2) + u(t),$$

for the unknown function values. The errors $u(t)$ are i.i.d. $N(0, \tau^2)$-variables, where τ^2 plays the role of an (inverse) smoothing parameter allowing for larger or enforcing smaller deviations in the development of $f(t)$. Assuming diffuse priors for initial values, i.e.,

$$p(f(1)) \propto \text{constant}$$

in case of first order random walks and

$$p(f(1)) \propto \text{constant}, \quad p(f(2)) \propto \text{constant}$$

in case of second order random walks, the joint prior for the vector $f = (f(1), \ldots, f(n))'$ is multivariate Gaussian with density

$$p(f) \propto \exp\left(-\frac{1}{2\tau^2} f' K_d f\right) \tag{10}$$

with precision matrix K_d given as in (7). Note that the random walk smoothness priors are partially improper since K_d has rank $n - d$. It can easily be shown that the posterior

$$p(f|y) \propto p(y|f)p(f) \tag{11}$$

is Gaussian with posterior mean

$$\hat{f} = E(f|y) = (I + \lambda K_d)^{-1} y, \tag{12}$$

where the smoothing parameter $\lambda = \sigma^2/\tau^2$ is defined as the noise-to-signal ratio, i.e., the ratio of error variance and variance of the random walk. The posterior covariance matrix is given by

$$\mathrm{Cov}(f|y) = \sigma^2 (I + \lambda K_d)^{-1}.$$

Thus, the Bayesian posterior mean estimate and the frequentist PLS-estimate coincide but the covariance matrices differ. To be more specific, the Bayesian posterior covariance matrix is larger (in terms of the Löwner order) than its frequentist counter part.

Since the posterior $p(f|y)$ is Gaussian, the posterior mean equals the posterior mode, which is the maximizer of the right-hand side in (11). Taking logarithms, it is straightforward to see that – up to a negative constant factor – the penalized (log-)likelihood criterion

$$l_{pen}(f) = \log p(y|f) + \log p(f)$$

is equal to the PLS criterion (6). This equivalence remains valid if we assume that errors are correlated so that the observation model is altered to

$$\varepsilon \sim N(0, \tau^2 R(\alpha)), \qquad y|f \sim N(f, \sigma^2 R(\alpha))$$

with (nonsingular) covariance matrix $R(\alpha)$, where α parameterizes the correlation structure. For example, the stochastic error process generating the correlation matrix might be a stationary autoregressive process of first or second order, i.e.,

$$\begin{aligned}
&\varepsilon(t) = \alpha\varepsilon(t-1) + u(t), \quad |\alpha| < 1, \\
&\varepsilon(t) = \alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + u(t), \ |\alpha_2| < 1, |\alpha_1| < 1 + \alpha_2
\end{aligned}$$

with i.i.d. Gaussian variables $u(t) \sim N(0, \sigma^2)$. The limiting cases $\alpha \to 1$ and $\alpha_1 \to 2$, $\alpha_2 \to -1$ lead to the (nonstationary) random walk models $RW(1)$ and $RW(2)$, respectively. Defining suitable distributions for the starting values, it can be shown that

$$\varepsilon \sim N(0, \sigma^2 K_{d,\alpha}^{-1})$$

with (nonsingular) precision matrices

$$K_{1,\alpha} = \begin{pmatrix} 1 & -\alpha & & & \\ -\alpha & 1+\alpha^2 & -\alpha & & \\ & \ddots & \ddots & \ddots & \\ & & -\alpha & 1+\alpha^2 & -\alpha \\ & & & -\alpha & 1 \end{pmatrix}$$

and

$$K_{2,\alpha} = \begin{pmatrix} 1 & -\alpha_1 & -\alpha_2 & & \dots \\ -\alpha_1 & 1+\alpha_1^2 & -\alpha_1(1-\alpha_2) & -\alpha_2 & \dots \\ -\alpha_2 & -\alpha_1(1-\alpha_2) & 1+\alpha_1^2+\alpha_2^2 & -\alpha_1(1-\alpha_2) & -\alpha_2 \dots \\ & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

$$\begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \\ \dots & -\alpha_2 & -\alpha_1(1-\alpha_2) & 1+\alpha_1^2+\alpha_2^2 & -\alpha_1(1-\alpha_2) & -\alpha_2 \\ \dots & & -\alpha_2 & -\alpha_1(1-\alpha_2) & 1+\alpha_1^2 & -\alpha_1 \\ \dots & & & -\alpha_2 & -\alpha_1 & 1 \end{pmatrix}.$$

In the limiting cases we obtain

$$\lim_{\alpha \to 1} K_{1,\alpha} = K_1, \qquad \lim_{\alpha_1 \to 2, \alpha_2 \to -1} K_{2,\alpha} = K_2,$$

i.e., the precision matrices of the corresponding random walks. For simplicity, we only take a closer look at $AR(1)$-processes ε and $RW(1)$-priors for f. Then, the PLS criterion (6) is replaced by the penalized weighted least squares (PWLS) criterion

$$\text{PWLS}(f) = (y-f)' K_\alpha (y-f) + \lambda f' K_1 f. \tag{13}$$

The PWLS estimate is then given by

$$\hat{f}_\alpha = (K_\alpha + \lambda K_1)^{-1} K_\alpha y. \tag{14}$$

The corresponding Bayesian hierarchical model is now

$$y|f \sim N(f, \sigma^2 K_\alpha^{-1}),$$

with the same Gaussian smoothness prior for f as in (10), with precision matrix K_1. The posterior $p(f|y)$ is Gaussian, but now with posterior mean

$$E(f|y) = \hat{f}_\alpha = (K_\alpha + \lambda K_1)^{-1} K_\alpha y,$$

so that the equivalence of the frequentist and Bayesian point estimate still holds. For $\alpha = 0$, $\hat{f}_\alpha$ reduces to the unweighted PLS estimate (8). For α close to 1, we may expect identification problems, since in the limiting case $\alpha \to 1$, we get

$$P_\alpha := K_\alpha + \lambda K_1 \to (1 + \lambda) K_1,$$

where K_1 is singular. These problems are reflected in the condition number

$$\kappa_\alpha = \frac{\lambda_{\max}(P_\alpha)}{\lambda_{\min}(P_\alpha)},$$

where $\lambda_{\max}(P_\alpha)$ and $\lambda_{\min}(P_\alpha)$ denote the largest and the smallest eigenvalue of P_α, respectively. Note that P_α is also the Bayesian posterior precision matrix of $\hat{f}$ thereby providing a measure for the variability of the estimate.

For large κ_α, inversion of P_α suffers from numerical instability as exemplified in Figure 1 for $n = 100$ time points. For increasing values of the autoregressive parameter α, the condition dramatically increases regardless of the value of the smoothing parameter. Small values of λ somewhat lower the effect, since the influence of K_1 on P_α is reduced, but qualitatively the effect remains the same. Note also, that the condition has been log-transformed in Figure 1 to enhance visibility. Hence, the value 10, for example, corresponds to a condition number of $\kappa_\alpha \approx 22000$.

The large condition number for values of α close to one reveals that the nonparametric function f is not well separable from the correlation and that, in particular, increasing variability of $\hat{f}$ is observed for $\alpha \to 1$. However, it seems plausible that we might still obtain a reasonable point prediction for the response vector y. To investigate this conjecture more closely, let us take a closer look at the behavior of the hat matrix P_α^{-1} projecting y on $\hat{y}$ in the limiting case $\alpha \to 1$.

Therefore we rewrite P_α^{-1} using the matrix inversion lemma Toutenburg (2003, Theorem A.18) as

$$(K_\alpha + D_1' \lambda I D_1)^{-1} = K_\alpha^{-1} - K_\alpha^{-1} D_1' \left(\frac{1}{\lambda} I + D_1 K_\alpha^{-1} D_1' \right)^{-1} D_1 K_\alpha^{-1}.$$

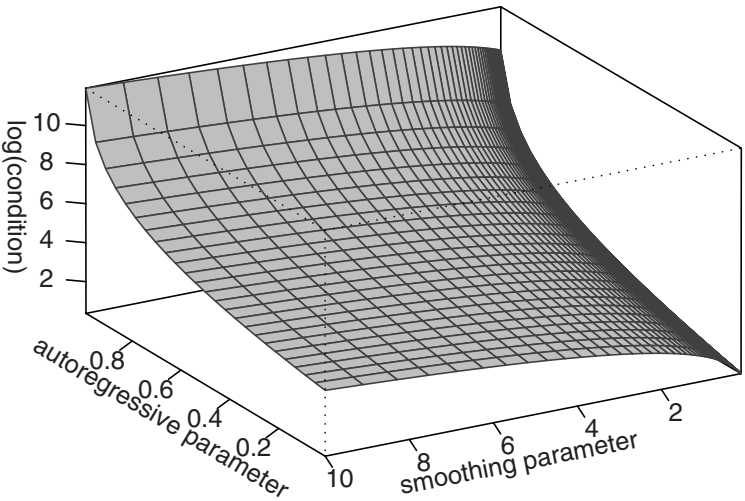

Fig. 1. Condition number κ_α for varying values of the autoregressive parameter α and the smoothing parameter λ when the nonparametric effect is modeled as first order random walk.

For $\alpha < 1$, the matrix K_α is regular and its inverse is given by

$$K_\alpha^{-1} = \frac{1}{1-\alpha^2} \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n-1} \\ \alpha & 1 & \alpha & & \alpha^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \alpha^{n-2} & & \ddots & \ddots & \alpha \\ \alpha^{n-1} & \alpha^{n-2} & \dots & \alpha & 1 \end{pmatrix}.$$

Straightforward calculations lead to the following expression for $\hat{y}$ in the limiting case $\alpha \to 1$:

$$\hat{y} = (K_\alpha + \lambda K_1)^{-1} K_\alpha y \underset{\alpha \to 1}{\longrightarrow} \begin{pmatrix} 1-\eta & 0 & \dots\dots & 0 & \eta \\ \eta & 1-2\eta & 0 \dots & 0 & \eta \\ \vdots & & \ddots & & \vdots \\ \eta & 0 & \dots & 1-2\eta & \eta \\ \eta & 0 & \dots & 0 & 1-\eta \end{pmatrix} y,$$

i.e.,

$$\hat{y}(t) = \begin{cases} (1-\eta)y(1) + \eta y(n) & t = 1 \\ \eta y(1) + (1-2\eta)y(t) + \eta y(n) & 2 \le t \le n-1 \\ \eta y(1) + (1-\eta)y(n) & t = n, \end{cases}$$

where $\eta = 0.5\lambda/(1+\lambda)$. Therefore the prediction for $\hat{y}(t)$ is always a weighted average of $y(1)$, $y(t)$ and $y(n)$, with the influence of $y(t)$ depending on the smoothness parameter. For $\lambda \to \infty$ (and $\eta \to 0.5$ correspondingly), the influence of $y(t)$ disappears and the overall prediction is just the constant $0.5(y(1) + y(n))$. In the contrary extreme ($\lambda \to 0$ or $\eta \to 0$) the prediction simply interpolates the observed time series.

These considerations lead to the following interpretation: If we try to estimate the trend while simultaneously accounting for correlation, serious multicollinearity problems arise if $\alpha \to 1$ since both the error term and the trend function follow the same stochastic structure. The prediction $\hat{y}$ is still well-behaved as a point estimate with meaningful limiting cases as the smoothing parameter λ is varied. However, the variability of both the estimate $\hat{f}$ and therefore the prediction $\hat{y}$ dramatically increases when $\alpha \to 1$.

It seems that the multicollinearity problem arises because the random walk smoothness prior for f and the stochastic process (10) for ε become so similar with α approaching 1. We may expect less problems with other priors for the trend which imply additional smoothness properties, e.g. for (Bayesian) penalized spline regression. Then we assume that $f(t)$ is (approximated as) a linear combination

$$f(t) = \sum_{j=1}^{p} \beta_j B_j^l(t)$$

of B-splines of degree l, defined for an equidistant grid of knots on the time axis. The vector f of function values can then be expressed as $f = X\beta$, where the design matrix X has elements $X[t, j] = B_j^l(t)$, $t = 1, \ldots, n$, and $j = 1, \ldots, p$. To enforce smoothness, the B-spline coefficient vector β obeys the same difference penalties or – in the Bayesian version – random walk priors as before. A standard choice are cubic B-splines and a $RW(2)$-prior. Then the observation model is

$$y \sim N(X\beta, \sigma^2 K_{d,\alpha}^{-1})$$

and the smoothness prior is Gaussian and of the form (10) again. As before, the PWLS estimate and the posterior mean estimate coincide and are given as

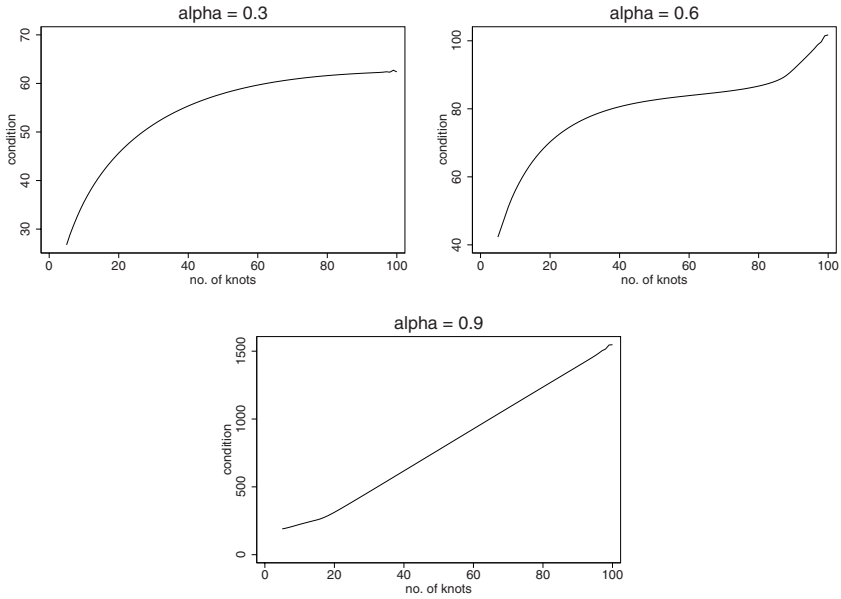

Fig. 2. Condition number κ_α for varying numbers of knots when the nonparametric effect is modeled as a cubic P-spline with second order difference penalty.

$$\hat{\beta}_\alpha = (X' K_{d,\alpha} X + \lambda K_d)^{-1} K_{d,\alpha} y.$$

For the popular choice of cubic B-splines and a $RW(2)$-prior for β, Figure 2 shows the condition number κ_α of the matrix $P_\alpha = X' K_\alpha X + K_2$ as a function of the number of knots and for different values of α. Although the shape of the condition number is quite different depending on the amount of correlation, all curves show the same qualitative behavior of an increasing condition number for larger numbers of knots. Note also the different scaling of the graphics: For high autoregressive correlation, the increase is much more dramatic than for moderate and small correlation.

We now move on a bit further and consider observation models of the form

$$y(t) = f(t) + \varepsilon(t) + \delta(t), \quad t = 1, \ldots, n \tag{15}$$

where, from a frequentist point of view, $f(t)$ is a (deterministic) trend function, $\varepsilon(t)$ is a (stationary) stochastic process inducing temporal correlation as before, and $\delta(t)$ are additional i.i.d. errors representing

pure measurement noise. Models of this form are the time series version of geostatistic ("kriging") models considered in the next section and allow for the estimation (or prediction) of both the trend function and the correlated error component. Assuming Gaussian errors, we have

$$y = f + \varepsilon + \delta, \quad \delta \sim N(0, \omega^2 I)$$

in matrix notation. As before, we adopt a basis function approach and approximate the trend through $f = X\beta$, while ε follows an $AR(1)$ or $AR(2)$-process with $\mathrm{Cov}(\varepsilon) = \sigma^2 K_{d,\alpha}^{-1}$. If the primary interest is in estimating f, inference will be based on the *marginal* distribution

$$y|f \sim N(X\beta, \sigma^2 V_\alpha^{-1})$$

where ε and δ are assumed to be independent, and

$$\sigma^2 V_\alpha^{-1} = \sigma^2 (K_\alpha^{-1} + \eta I), \quad \eta = \omega^2/\sigma^2.$$

is the covariance matrix of $\varepsilon + \delta$.

The resulting PWLS estimate for β is

$$\hat{\beta}_\alpha = (X' V_\alpha X + \lambda K_d)^{-1} V_\alpha y.$$

In the special case of B-splines of degree zero, corresponding to random walk models, we have $X = I$ and $f = \beta$, and it may again be interesting to take a closer look at

$$\hat{f}_\alpha = (V_\alpha + \lambda_1 K_1)^{-1} V_\alpha y$$

in the limiting case $\alpha \to 1$.

Problems of (weak) identifiability become quite obvious from a Bayesian perspective if we consider the *conditional* distribution of y, given the trend f and the stochastic process ε generating correlation, i.e.,

$$y|f, \varepsilon \sim N(f + \varepsilon, \sigma^2 I),$$

with $f = X\beta$. This means, we attempt to separate observation y into three components f, ε and δ differing only through their prior specifications. If the smoothness prior for f and the stochastic process prior for ε have similar stochastic structure, then it will obviously be difficult to distinguish them given a finite sample of data y. Moreover, the Bayesian interpretation also reveals, that trend and correlation are per se connected quite closely. If long range correlation is present in the data (corresponding to $\alpha \approx 1$ in the $AR(1)$ example), these correlation

will almost appear as a smooth trend in the data. Vice versa, wiggly trends may be equivalently interpreted as some kind of shorter ranged correlation. Compare also the simulation result at the very end of this section.

Obviously, random walk models for f in combination with a stationary autoregressive process, where the parameters approach the boundary of the stationary region, are a simple prototype for weak identifiability. It is easily derived that the posterior $p(f, \varepsilon|y)$ is Gaussian, and the posterior mean estimates $\hat{f}$, $\hat{\varepsilon}$ satisfies

$$\begin{pmatrix} I + \lambda_1 K_d & I \\ I & I + \lambda_2 K_{d,\alpha} \end{pmatrix} \begin{pmatrix} \hat{f} \\ \hat{\varepsilon} \end{pmatrix} = \begin{pmatrix} y \\ y \end{pmatrix},$$

where $\lambda_1 = \omega^2/\tau^2$, $\lambda_2 = \omega^2/\sigma^2$. In the limiting case $\alpha \to 1$ the matrix becomes singular and an identifiability problem arises if we want to separate f from ε. In the following we explore these (weak) identifiability issues empirically through some simulation experiments, focussing on the situation where time series data are generated from models of the form (15), with

$$y(t) = \sin(t) + \varepsilon(t) + \delta(t), \quad t = 1, \ldots, 100,$$

and grid length $\Delta t = 0.25$. We set $\sigma^2 = \text{Var}(\varepsilon(t)) = \omega^2 = \text{Var}(\delta(t)) = 0.5$, which is close to the empirical variance of the sine function, so that all three components have about the same variability. All data sets were generated for $\alpha = 0.3$ (low correlation), $\alpha = 0.6$ (medium correlation), and $\alpha = 0.9$ (strong correlation).

For estimation, the true trend was approximated through penalized P-splines, varying the degree l of the spline functions, the smoothness penalty ($RW(1)$ or $RW(2)$) for B-spline coefficients, and the number of knots. For each selected combination of α-values and B-spline tuning parameters, 50 data sets were generated according to the specific model. The models were fitted either with full Bayesian inference using MCMC or empirical Bayesian inference using mixed model technology. These inference techniques are described in Fahrmeir, Kneib and Lang (2004) and Lang and Brezger (2004), and are implemented in the software BayesX (Brezger, Kneib and Lang (2007)). For each data set, goodness of fit was assessed through

$$SQ(f) = \sum_{t=1}^{100} (f(t) - \hat{f}(t))^2,$$
$$SQ(\varepsilon) = \sum_{t=1}^{100} (\varepsilon(t) - \hat{\varepsilon}(t))^2,$$
$$SQ(y) = \sum_{t=1}^{100} (y(t) - \hat{y}(t))^2,$$

and variability measured through

$$VS(f) = \sum_{t=1}^{100} \mathrm{Var}(\hat{f}(t)),$$
$$VS(\varepsilon) = \sum_{t=1}^{100} \mathrm{Var}(\hat{\varepsilon}(t)),$$
$$VS(y) = \sum_{t=1}^{100} \mathrm{Var}(\hat{y}(t)).$$

Figures 3–6 display boxplots of these characteristics, resulting from estimation with different combinations of α and B-splines for the 50 data sets, respectively.

These figures and additional ones in Eschrich (2007) provide the following empirical evidence:

- With increasing correlation, quality of estimation of the components f and ε decreases (Figure 3, c,e). In contrast, the predictions for the response $\hat{y} = \hat{f} + \hat{\varepsilon}$ for the sum remain comparably stable regardless of the amount of correlation (Figure 3, a). This confirms the results for increasing α that we discussed from a theoretical perspective earlier in this section: While separation between f and ε proves to be difficult, the overall fit remains well identified. Note that even the variance of $\hat{y}$ is relatively stable while variability of both $\hat{f}$ and $\hat{\varepsilon}$ increases.
- Figure 4 investigates the dependence of the results on the number of knots for the standard choice of cubic P-splines with $RW(2)$-priors, corresponding to a rather smooth prior. In this case, results seem to be rather insensitive to the number of knots, as opposed to what might have been guessed from the condition number displayed in Figure 2. For the goodness of fit measures $SQ(f)$ and $SQ(\varepsilon)$, we even obtain improved results for an increased number of knots and therefore a better separation of trend and correlation.

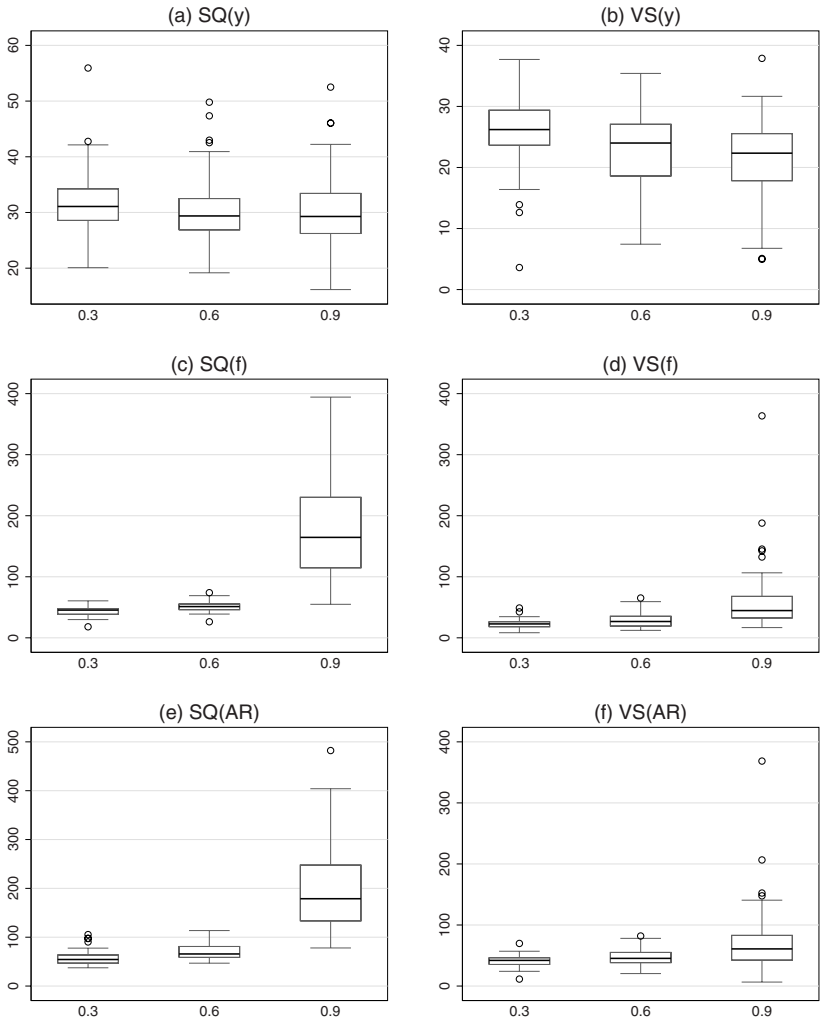

Fig. 3. Goodness of fit and variability measures for varying values of the autoregressive parameter α. The nonparametric effect is modeled as a cubic P-spline with second order random walk penalty and 40 knots.

- In contrast, if the prior for the nonparametric trend does not enforce smoothness but is closer to the $AR(1)$-process results are qualitatively different. Figures 5 and 6 show results for zero degree P-splines and a high amount of correlation for the autoregressive component. When varying the number of knots (Figure 5), both the fit of the nonparametric effect and the autoregressive component wors-

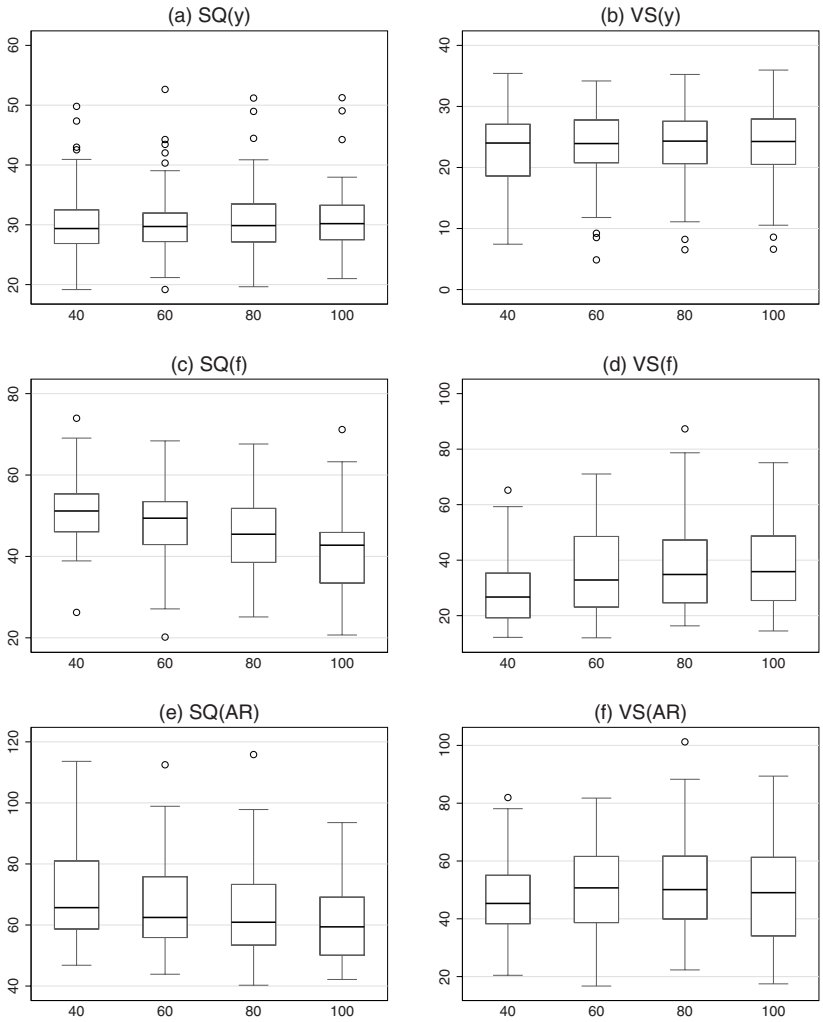

Fig. 4. Goodness of fit and variability measures for varying numbers of knots for the nonparametric effect. The autoregressive parameter is fixed at $\alpha = 0.6$ and the nonparametric effect is modeled as a cubic P-spline with second order random walk penalty.

ens, while the overall fit remains roughly the same. When comparing $RW(1)$ and $RW(2)$ priors for the nonparametric effect, identification somewhat worsens for the first order random walk, which is closer to the $AR(1)$-process than the $RW(2)$ prior (Figure 6). Overall, as expected, the worst choice in terms of identifiability is a zero

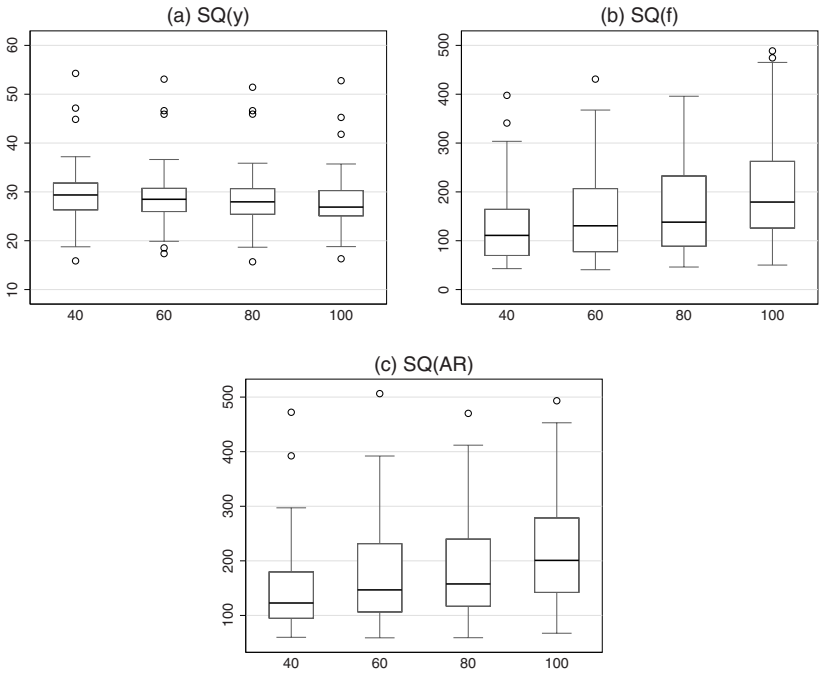

Fig. 5. Goodness of fit and variability measures for varying numbers of knots for the nonparametric effect. The autoregressive parameter is fixed at $\alpha = 0.9$ and the nonparametric effect is modeled as a piecewise constant P-spline with first order random walk penalty.

degree P-spline with random walk of first order as smoothness prior for the trend and a large number of knots.

For the results presented so far, both the autoregressive error $\varepsilon(t)$ and the independent error $\delta(t)$ have been generated anew in each simulation run. To be able to derive mean estimates averaged over the simulation runs, we repeated parts of the simulations with a fixed sequence of autoregressive errors (but still with varying independent errors, of course). Figure 7 shows one exemplary result from these simulations, where the nonparametric effect is modeled as a cubic P-spline with 20 knots and second order random walk penalty. The autocorrelation parameter is fixed at the high value, such that the generated autocorrelated error varies relatively slowly as time progresses. Therefore a large fraction of the autoregressive process is absorbed by the nonparametric effect and the original sine curve as well as the autoregressive

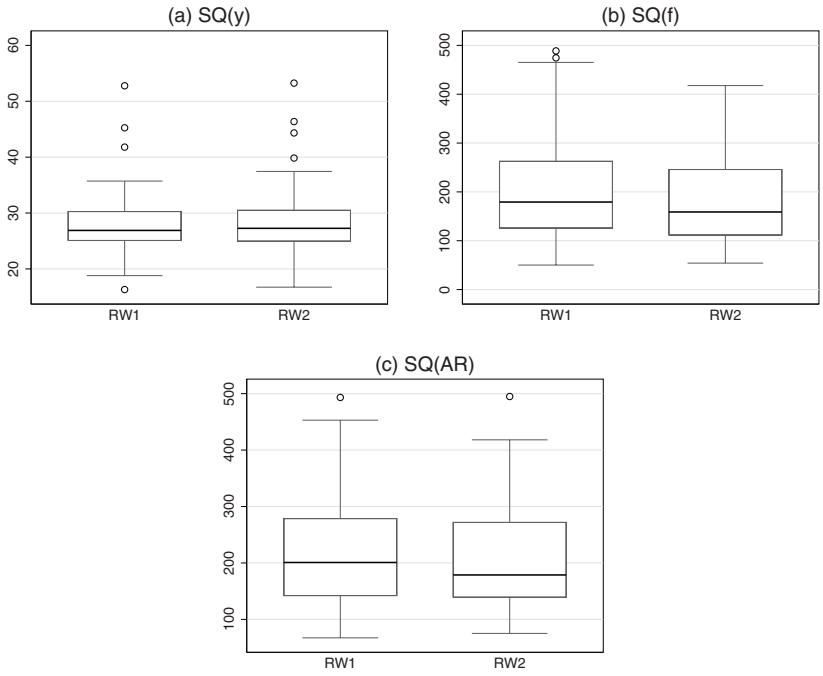

Fig. 6. Goodness of fit and variability measures for varying specifications of the prior for the nonparametric effect. The autoregressive parameter is fixed at $\alpha = 0.9$ and the nonparametric effect is modeled as a piecewise constant P-spline with 100 knots.

component are not very well identified. This again indicates, that trend estimation and modelling of correlation are not opponent concepts but overlapping areas of statistical inference.

3 Spatial Correlation

The modelling approaches for nonparametric trend estimation and temporal correlation considered in the previous section can be extended to estimation of spatial surfaces while simultaneously taking into account spatial correlation. Therefore we replace the univariate temporal model (15) with the bivariate spatial model

$$y(s) = f(s) + \varepsilon(s) + \delta(s)$$

where $s = (s_x, s_y) \in S \subset \mathbb{R}^2$ represents continuous coordinates in some suitable spatial region S, $f(s)$ models a smooth spatial trend