Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 131

Instant Trend-Seasonal Decomposition of Time Series with Splines

Luis Francisco Rosales, Tatyana Krivobokova

November 2012

Wilhelm-Weber-Str. 2 · 37073 Goettingen · Germany
Phone: +49-(0)551-3914066 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

Instant Trend-Seasonal Decomposition of Time Series with Splines

Luis Francisco Rosales[1]
Tatyana Krivobokova
Georg-August-Universität Göttingen

15th November 2012

Abstract

We present a nonparametric method to decompose a times series into trend, seasonal and remainder components. This fully data-driven technique is based on penalized splines and makes an explicit characterization of the varying seasonality and the correlation in the remainder. The procedure takes advantage of the mixed model representation of penalized splines that allows for the simultaneous estimation of all model parameters from the corresponding likelihood. Simulation studies and three data examples illustrate the effectiveness of the approach.
Key Words: Penalized splines, Mixed model, Varying coefficient, Correlated remainder.

1 Introduction

There are many approaches in the literature to the problem of trend extraction from a time series. The interested reader can refer to Alexandrov et al. (2012) or Pollock (2006) for a detailed discussion on the history of the problem, and the advantages and disadvantages of the different alternatives available to treat it. In general, detrending methods can be grouped in one of the following classes: the model based approach, nonparametric filtering, singular spectrum analysis and wavelets. The model based approach is the *de facto* tool in economics, even though it requires a specification of the structure of the time series under analysis by either an ARIMA or an state-space model (see e.g.

[1]Courant Research Center "Poverty, equity and growth" and Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Wilhelm-Weber-Str. 2, 37073 Göttingen, Germany

Findley et al., 1998; Maravall and Caporello, 2004). Non-parametric methods, in contrast, do not require such specification and are popular because of their simplicity. Among the most cited approaches are the filters proposed by Henderson (1916), Hodrick and Prescott (1997) and Cleveland et al. (1990). Singular spectrum analysis and Wavelets are most commonly used in geosciences, but their applications in economics have increased in recent years (see e.g. Pollock and Cascio, 2007; Ramsay and Lampart, 1998).

In general, all nonparametric techniques used for detrending of time series have two main drawbacks: the remainder structure is not modelled explicitly (and is left unspecified) and the smoothing parameter choice is done in a not data-driven way, which ultimatly can lead to a wrong decomposition. The method we develop in this article, allows for the simultaneous decomposition of a time series into a smooth overall trend, a seasonal component (which can vary over time) and a remainder component, modeled as an $ARMA(p, q)$ process. Thereby, all model components, including smoothing and covariance matrix parameters are estimated in a single run, optimizing an appropriate likelihood function. We investigate the properties of our detrending technique in a simulations study and present several real-data examples.

The paper is organized as follows: in Section 2 we introduce the method and propose a procedure to select the input parameters for the model. Section 3 shows a Monte Carlo simulation study that compares the results of our method with the STL procedure developed by Cleveland et al. (1990). The application of the method is then illustrated by data examples in Section 4. Section 5 presents some conclusions and closes the paper.

2 Decomposition of time series with splines

Consider a smooth decomposition scheme of a time series

$$y(t_i) = \tau(t_i) + \varsigma(t_i) + \epsilon(t_i), \quad i = 1, \ldots, n, \tag{1}$$

where $\tau(t)$ is an unknown smooth function that represents an overall trend, $\varsigma(t)$ is a function that describes seasonal fluctuations and $\epsilon(t)$ is an $ARMA(p,q)$ process. The seasonal component will be modelled as $\varsigma(t_i) = \alpha(t_i)\cos(t_i\omega) + \gamma(t_i)\sin(t_i\omega)$ where $\alpha(t)$ and $\gamma(t)$ are unknown smooth functions that modulate the seasonal pattern over time (varying coefficients). In principle, $\varsigma(t)$ can easily be generalized to include components with different frequencies and corresponding varying coefficients. Subsequently, we show the estimation of (1) with $\varsigma(t_i) = \alpha(t_i)\cos(t_i\omega)$ only, stating that the extension to more components is immediate. Additionally we will refer to $f(t_i) = \tau(t_i) + \varsigma(t_i)$ as the smooth part of the decomposition.

Assuming that parameters p, q and ω are known, let map the time span into the unit interval $[0,1]$ as $x_i = t_i / \sum_{i=1}^n t_i$ and employ low-rank spline smoothing to estimate (1). That is, smooth functions τ and α are estimated by splines solving

$$\min_{s_\tau \in \mathcal{S}_\tau, s_\alpha \in \mathcal{S}_\alpha} \quad \left[\frac{1}{n\sigma^2} \sum_{i=1}^n \sum_{j=1}^n \{y(x_i) - s_\tau(x_i) - s_\varsigma(x_i)\} R_{ij}^{-1} \{y(x_j) - s_\tau(x_j) - s_\varsigma(x_j)\} \right.$$
$$\left. + \quad \lambda_\tau \int_0^1 \{s_\tau(x)^{(m_\tau)}\}^2 dx + \lambda_\alpha \int_0^1 \{s_\alpha(x)^{(m_\alpha)}\}^2 dx \right], \tag{2}$$

where $s_\varsigma(x) = s_\alpha(x)\cos(x\omega)$ and R_{ij}^{-1} is the (ij)th element in the inverse of covariance matrix $\boldsymbol{R}(\boldsymbol{\phi}) = \sigma^2 \operatorname{Cor}\{\epsilon(t_i), \epsilon(t_j)\}_{i,j=1}^n$, with a $(p+q)$-dimensional parameter $\boldsymbol{\phi}$ of an $ARMA(p,q)$ process. $\mathcal{S}_j$, $j \in \{\tau, \alpha\}$ denote spline spaces of degree $2m_j - 1$, which are based on k_j knots and hence consist of $2m_j - 2$ times differentiable functions, that are polynomials of degree $2m_j - 1$ between each two consecutive knots. To solve (2) we make the following assumptions.

(A1) $\tau \in W^{m_\tau}[0,1]$ and $\alpha \in W^{m_\alpha}[0,1]$, where $W^m[0,1]$ is the Sobolev space of functions with $m-1$ continuous derivatives and integrable mth derivative.

(A2) Both, observations and spline knots, are equidistant.

(A3) The numbers of knots k_j, $j \in \{\tau, \alpha\}$ satisfy $k_j = \text{const}\, n^{\nu_j}$, $\nu_j \in (1/(2m_j), 1)$, where

"const" denotes a generic positive constant.

Since $\mathcal{S}_j$ is a linear space of dimension $k_j + 2m_j$, each spline function s_τ, s_α can be represented as a linear combination of some basis functions of $\mathcal{S}_j$. The B-spline basis is popular due to its numerical stability and will be used in our implementation. Denote the basis matrix by $\boldsymbol{B}_j = \{\boldsymbol{B}_j(x_1)^t, \ldots, \boldsymbol{B}_j(x_n)\}^t$, where $\boldsymbol{B}_j(x)$ is a $(k_j + 2m_j)$-dimensional row vector of B-spline basis of degree $2m_j - 1$ based on k_j knots, $j \in \{\tau, \alpha\}$. Representing now $s_j(x) = \boldsymbol{B}_j(x)\boldsymbol{\theta}_j$, turn (2) into the minimization problem over $\boldsymbol{\theta}_\tau$ and $\boldsymbol{\theta}_\alpha$. It remains to estimate λ_τ, λ_α and $\boldsymbol{\phi}$. In principle, this can be done minimizing a version of the generalized cross validation (GCV) as a function of all these unknown parameters. However, it is well-known that the GCV criterion can be extremely unstable and the resulting parameter estimators have very large variance, see Krivobokova (2012). Therefore, we opt to make use of the mixed model representation of penalized splines (see e.g. Wand, 2003), which allows to estimate all the model parameters from the corresponding likelihood function and is known to be much more stable than GCV and similar methods.

Let decompose each component $\boldsymbol{B}_j\boldsymbol{\theta}_j = \boldsymbol{B}_j(\boldsymbol{F}_b^j\boldsymbol{b}_j + \boldsymbol{F}_u^j\boldsymbol{u}_j) = \boldsymbol{X}_j\boldsymbol{b}_j + \boldsymbol{Z}_j\boldsymbol{u}_j$, so that it holds $(\boldsymbol{F}_u^j)^t\boldsymbol{F}_b^j = (\boldsymbol{F}_b^j)^t\boldsymbol{D}_j\boldsymbol{F}_b^j = 0$ and $(\boldsymbol{F}_u^j)^t\boldsymbol{D}_j\boldsymbol{F}_u^j = \boldsymbol{I}_{k_j+m_j}$, where $\boldsymbol{D}_j$ is such that $\int_0^1\{\boldsymbol{B}_j(x)^{(m_j)}\boldsymbol{\theta}_j\}^2dx = \boldsymbol{\theta}_j^t\boldsymbol{D}_j\boldsymbol{\theta}_j$. This decomposition is not unique, since $\boldsymbol{D}_j$ is not of full rank, but we follow the strategy suggested in Wood (2006). Thus, the best linear unbiased predictor in the linear mixed model

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_\tau\boldsymbol{b}_\tau + \boldsymbol{Z}_\tau\boldsymbol{u}_\tau + \mathrm{diag}\{\cos(\boldsymbol{x}\omega)\}\{\boldsymbol{X}_\alpha\boldsymbol{b}_\alpha + \boldsymbol{Z}_\alpha\boldsymbol{u}_\alpha\} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}\left(\boldsymbol{0}_n, \sigma^2\boldsymbol{R}(\boldsymbol{\phi})\right), \quad \boldsymbol{u}_j \sim \mathcal{N}(\boldsymbol{0}_{k_j+m_j}, \sigma_{u_j}^2\boldsymbol{I}_{k_j+m_j}), \quad j \in \{\tau, \alpha\}, \end{aligned} \tag{3}$$

for $\boldsymbol{Y} = \{y(x_1)\ldots, y(x_n)\}^t$ and $\boldsymbol{\epsilon} = \{\epsilon(x_1), \ldots, \epsilon(x_n)\}^t$, equals to the solution of (2) with $\lambda_j = \sigma^2/\sigma_{u_j}^2$. From the likelihood function that corresponds to (3) one can estimate simultaneously $f(t)$ and all the covariance parameters, that is $\boldsymbol{\phi}$, σ^2, $\sigma_{u_\tau}^2$ and $\sigma_{u_\alpha}^2$, which is readily implemented in any standard statistical software for mixed models (e.g. function

`lme` in R).

To fit model (3), the number of knots k_j and the penalization order m_j need to be specified, as well as parameters p, q and ω, which have been assumed known so far. Regarding the number of knots k_j, assumption (A3) reflects the result obtained in (Claeskens et al., 2009) that, once enough knots are taken, increasing its number does not change the estimator in terms of the average mean squared error. Hence, in practice, one can choose the number of knots to be rather large ($k_\tau = k_\alpha$ between $20 - 150$, depending on the sample size), but it can still be much smaller than the number of observations. Due to the smoothness of the cosine function, the penalization order m_α for the varying coefficient is of minor importance and one can set it to the standard choice $m_\alpha = 2$. Thereafter, our main concern will be in the selection of m_τ, for which we set an initial value $m_\tau = m_\tau^*$ that helps us select all the other parameters, and whose value will be later updated making use the criterion proposed by Krivobokova (2012), which, once adjusted to our model, takes the form

$$m_\tau = \underset{m_\tau \in \mathbb{N}}{\operatorname{argmin}}\, \mathcal{R}(m_\tau) = \underset{m_\tau \in \mathbb{N}}{\operatorname{argmin}} \quad \left| \boldsymbol{Y}^t (\boldsymbol{I}_n - \boldsymbol{G}_\varsigma)(\boldsymbol{I}_n - \boldsymbol{S}_{\boldsymbol{\tau}}) \boldsymbol{S}_{\boldsymbol{\tau}}{}^2 (\boldsymbol{I}_n - \boldsymbol{G}_\varsigma) \boldsymbol{Y} - \sigma^2 \left\{ \operatorname{tr}(\boldsymbol{S}_{\boldsymbol{\tau}}^{\boldsymbol{2}}) - m_\tau \right\} \right|, \tag{4}$$

with $\boldsymbol{S}_\tau(\boldsymbol{R}) = \boldsymbol{C}_\tau(\boldsymbol{C}_\tau^t \boldsymbol{R}^{-1} \boldsymbol{C}_\tau + \boldsymbol{D}_\tau)^{-1} \boldsymbol{C}_\tau^t \boldsymbol{R}^{-1}$, $\boldsymbol{C}_\tau = [\boldsymbol{X}_\tau, \boldsymbol{Z}_\tau]$ and $\boldsymbol{D}_\tau = \sigma^2 / \sigma_{u_\tau}^2 \operatorname{diag}\{\boldsymbol{0}_{m_\tau}, \boldsymbol{1}_{k_\tau + m_\tau}\}$, and $\boldsymbol{G}_\varsigma$ as defined in the appendix.

To select the starting value m_τ^* we bring to bear the result of Krivobokova and Kauermann (2007) (see also Kauermann et al., 2011). It is well-known, that the smallest error structure misspecifications lead to a severely overfitted estimate of the regression function, if it is obtained with GCV or similar criteria, see Opsomer et al. (2001). In contrast, the estimator obtained from the mixed model representation of penalized splines is remarkably robust to misspecifications of the remainder's structure. That is, if the data follow (1),

but the model

$$\boldsymbol{Y} = \boldsymbol{X}_\tau \boldsymbol{b}_\tau + \boldsymbol{Z}_\tau \boldsymbol{u}_\tau + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}_n, \sigma^2 \tilde{\boldsymbol{R}}), \tag{5}$$

is computed, the resulting estimator of $\hat{\boldsymbol{\tau}} = \boldsymbol{S}_\tau(\tilde{\boldsymbol{R}})\boldsymbol{Y}$ under some working correlation $\tilde{\boldsymbol{R}}$ is barely affected as long as the penalty order m_τ is chosen appropriately. Hence, a possible strategy to select an initial value for m_τ could be through the visual inspection of the trend estimations resulting from (5). Alternatively, if we interpret the action of matrix $\boldsymbol{S}_\tau$ (depending on m_τ) over $\boldsymbol{Y}$ as a low-pass filter (see e.g. Hastie and Tibshirani, 1990), we could also select the penalty order inspecting the frequency domain of the filter's transfer function, see section 4 for more details.

With a starting value m_τ^*, we proceed to determine parameters p, q and ω which can be done by fitting an $ARMA(p, q)$ model for the residuals resulting from (5). With all the parameters selected we go back to criterion (4) and re-estimate m_τ. All together, we fit (1) by pursuing the following steps:

1. Fix $k_\tau = k_\alpha$ to be sufficiently large ($20 - 150$, depending on n).
2. Determine a working m_τ^* fitting (5) for different values of m_τ.
3. Estimate (5) under m_τ^* and obtain the residuals.
4. Fit an $ARMA(p, q)$ model with a seasonal component to the residuals of (5) to determine p, q and ω.
5. Obtain the optimal m_τ by applying the $\mathcal{R}(m_\tau)$ criterion given in (4).
6. Estimate (3) with obtained m_τ, $m_\alpha = 2$, $k_\tau = k_\alpha$, p, q and ω.

Of course, it can happen that the data have a more complicated structure, that can not be described by (1). This will be evident if the remainder structure turns out to be non-stationary, the estimated parameters of the $ARMA(p, q)$ model lie close to the boundary

of stationarity, or a single seasonal frequency is inadequate. In any of these cases, one can try to include more frequencies, and/or to work with the first differences of the original series.

3 Simulation study and comparison with STL

We generated $n = 500$ data points with the following setup for the decomposition scheme (1) under $x_i = t_i / \sum_{i=1}^{n} t_i$. The trend component is generated as $\tau(x_i) = 6\beta_{30,17}(x_i)/10 + 4\beta_{3,11}(x_i)/10$, with $\beta_{r,s}$ the beta function defined as $\beta_{r,s}(x_i) = \Gamma(r+s)\{\Gamma(r)\Gamma(s)\}^{s-1}$; the seasonal component follows $\varsigma(x_i) = \alpha_{1.9,0.9}(x_i)\cos(\omega x_i)$ where $\alpha_{u,v}(x_i) = (1/2\pi)\{1 + u^2 + v^2 + 2u(v-1)\cos(\pi(2x_i - 1)) - 2v\cos(2\pi(2x_i - 1))\}^{-1}$, with a period of 20 observations, i.e. $\omega = 2\pi(n/20)$; and for the remainder component we simulate a first order autoregressive process with autocorrelation coefficient equal to 0.4. Furthermore the trend, seasonal and remainder components are re-scaled so their variances are 1, 0.5 and 0.1 respectively to assure a reasonable signal-to-noise ratio.

Figure 1 shows a simulated time series with the setting previously detailed. To fit (1) for this particular case we follow the strategy given in section 2. Namely, for $m_\alpha = 2$ and $k_\tau = k_\alpha = 50$, we start by selecting m_τ^* by inspecting the fits of (5) under different penalization choices. If the data were circular, the operation $\hat{\boldsymbol{\tau}} = \boldsymbol{S}_\tau(\tilde{\boldsymbol{R}})\boldsymbol{Y}$ obtained after fitting (5) could be seen as a stationary invariant linear filter with a unique impulse-response function $\kappa(t)$ centered at the main diagonal of $\boldsymbol{S}_\tau(\tilde{\boldsymbol{R}})$. The effect of this matrix on input $\boldsymbol{Y}$ is usually explored by taking the Fourier transform of the impulse-response function, i.e. $K(\omega) = \sum_{j=-\infty}^{\infty} \exp\{-i\omega t_j\}\kappa(t_j)$, also called transfer function. Figure 2 shows the impulse-response function, its Fourier transform and the resulting estimated trend for two choices of m_τ when fitting (5). The gray line in panel (b) indicates the frequency of the seasonal component in the generated data, and shows that for $m_\tau = 2$ the frequency response function of the trend component lie in a range of frequencies that

covers the one characterizing the seasonal component, and hence captures both parts indistinguishably. The selection of $m_\tau = 5$ on the other hand, seems more adequate and it is what we chose in this example.

With the starting value $m_\tau^* = 5$ selected, we use the residuals in (5) to identify $\omega = 20$, $p = 1$ and $q = 0$. We then go back to (4) and verify that indeed $m_\tau = 5$ is the optimal choice. All together, the estimation of (3) is set up with the following parameters: $m_\tau = 5,\ m_\alpha = 2,\ k_\tau = k_\alpha = 50,\ p = 1,\ q = 0,\ \omega = 20$.

We continue by comparing the performance of the splines decomposition approach with a benchmark alternative commonly used by practitioners, namely the STL procedure developed by Cleveland et al. (1990), which is another filtering method for decomposing a time series into trend, seasonal and remainder components consisting on systematic applications of LOWESS (locally weighted scatterplot smoother). LOWESS was presented by Cleveland (1979) as a smoothing method that performs a polynomial fit of degree d for a response y on a locally weighted version of covariate x. The weighting function is defined by a span parameter h as $w_i(x_0) = W(|x_i - x_0|/\delta_h(x_0))$ where $\delta_h(x_0)$ is the distance of the hth furthest x_i from x_0 and $W(l) = (1 - l^3)^3$ for $l \in [0, 1)$ and 0 otherwise.

The STL procedure is then built from systematic applications of the LOWESS smoother embedded in two loops: an inner loop that performs a seasonal smoothing that updates the seasonal component followed by a trend smoothing that updates the trend component; and an outer loop executed for robustness.

Consider for example the case of monthly data with yearly seasonality. In thi scase the updates at the $(j + 1)$th pass of the inner loop would be computed in the following way: i) detrend the series with the jth update of the trend component by $y(x_i) - \tau^j(x_i)$; ii) build a set of cycle-subseries based on the detrended series by grouping all Januaries, all Februaries, etc., smooth them with LOWESS, and build a (temporary) seasonal series $c^{j+1}(x_i)$; iii) construct a low-pass filter of the (temporary) seasonal series $L^{j+1}(x_i)$;

iv) update the seasonal series detrending $c^{j+1}(x_i)$ by $\varsigma^{j+1}(x_i) = c^{j+1}(x_i) - L^{j+1}(x_i)$; v) deseasonalize the original time series with $y(x_i) - \varsigma^{j+1}(x_i)$; and vi) update the trend component by smoothing the resulting deseasonalized series by LOWESS and obtain $\tau^{j+1}(x_i)$. With respect to the outer loop, its updates are performed for robustness and operate by modifying the weighting functions in steps ii) and vi) of the inner loop so the effect of aberrant observations in the data (measured by the local magnitude of the remainder) is diminished.

At this point it is convenient to highlight that both the STL and the splines approach are similar in various aspects. Most importantly, if we were to follow the circularity assumption of the data for interpretation purposes (see e.g. Cleveland et al., 1990), both methods could be seen as stationary symmetric linear filters for each of its components modulated by different weighting (or impulse-response) functions. In addition, the two of them can handle missing values, and both require the setting of various constants in order to be implemented. However, there are also some differences that must be noticed. An illustrative comparison between the STL and spline method to filter the simulated time series is shown in Figure 3. For the splines method we set up the model with $m_\tau = 5,\ m_\alpha = 2,\ k_\tau = k_\alpha = 50,\ p = 1,\ q = 0,\ \omega = 20$, as indicated at the beginning of this section, and for the STL procedure we consider $d_\tau = 1,\ h_\tau = 39$ for the trend component and $d_\varsigma = 1,\ h_\varsigma = 7$ for the seasonal component. As it can be seen in panel (a) of Figure 3, both methods produce very similar results for the trend component, up to certain wiggliness in the STL case. We do not explore the performance of STL for polynomials of degree greater than 1, and conclude that the differences between both methods are small in magnitude.

With respect to the seasonal component, the differences between both approaches are more notorious. Even though both procedures allow for the variation of the seasonal part across time, the behavior of the STL fit seems to be much more variable. To understand

the difference in the shape of both seasonal components we start by noting that while the spline method performs smoothing for the data series along the index x, the STL alternative does it for each cycle-subseries according to the span chosen for the LOWESS window. Clearly, once the smoothed subseries are re-arranged according to the original time sequence, a rough path can be observed. Furthermore, in this example we chose a span window of size $h_\varsigma = 7$ (a very small value when compared to the sample size $n = 500$), which produces a high variance estimation for this component (with small bias). We based our selection on the inspection of the seasonal-diagnostic plot under different scenarios of h_ς, as suggested in Cleveland et al. (1990). In fact, to our knowledge, there is no data-driven method for the selection of any of the STL parameters, including the crucial smoothing parameter h_ς. Consequently, the seasonal component could be undersmoothed, as it happens in this example.

Lastly, regarding the scatterplots of the remainders in panel (c) of Figure 3, we note that the remainder part from the spline approach (black circles) follow better the true remainder than those of the STL (gray circles), which is an obvious result of the better performance of the spline method to fit the seasonal component, and, consequently, allows for a more accurate characterization of an ARMA model for the remainder. Furthermore, we stress the fact that in the splines procedure the estimation of the model for the remainder part is performed simultaneously with the rest of parameters, as presented in (3), and not in a follow up computation as would be in the STL case. The last panel of Figure 3 compares the estimations for the two methods of the trend and seasonal components when plotted together.

We close this section by reporting the performance of the penalized splines method when compared to the STL alternative in a Monte Carlo simulation study considering 500 realizations of the remainder component. The results are presented in terms of the component-wise average mean squared error (AMSE) of each method and are illustrated in Figure 4.

As expected from the example presented in Figure 3, the performance of both methods is quite similar for the estimation of the trend with a slightly higher bias for the STL. However, when it comes to comparing the seasonal parts, a clear superiority of the splines procedure is revealed, showing to be less variable and more accurate than its counterpart.

4 Applications

In this section we present three examples in different resolutions to illustrate the application of the method described in section 2. The first data set consists of hourly temperature in California; the second one addresses weekly stock prices in Peru; and the the third one is based on monthly prices of pork meat in the European Union. Based on a previous inspection of the distribution of the residuals we consider appropriate to use a log transformation of each time series, implying an underlying multiplicative structure in decomposition (1).

4.1 Hourly Temperature in California

The source of the data is the United States' National Climatic Data Center (NOAA) and comprehends hourly averages of temperature in California measured in Fahrenheit degrees. The sample covers the period 01/01/2010 – 01/31/2010, with a total of 744 observations.

To set up the model we start by fixing the number of knots to $k_\tau = k_\alpha = 50$. For the selection of the other parameters we apply the strategy presented in section 2. A classical spectral analysis on the residuals after fitting (5) with a starting value $m_\tau^* = 4$ showed the need for the inclusion of i) frequencies $\omega_1 = 2\pi(744/24)$, $\omega_2 = 2\pi(744/12)$ and $\omega_3 = 2\pi(744/6)$; and ii) a second order autocorrelation structure AR(2) to characterize the remainder component. Lastly, the $\mathcal{R}(m_\tau)$ criterion selected a penalization parameter $m_\tau = 3$ for the trend component.

Figure 5 presents the decomposition of the temperature series during January 2010. The

upper-most panel presents the data (gray line), the fitted trend component (black line) and the addition the fitted trend and seasonal parts (dashed lines). As it can be seen, the features of the original time series are replicated by the model, as the trend shows a mildly positive slope (upper-most panel), and the magnitude of the seasonality seems to be enlarged with time (middle panel). These features are expected since historically the intraday temperature range is positively correlated with its level.

The two plots at the bottom panel of Figure 5 show that the selected characterization of the remainder by an AR(2) is indeed appropriate. The gray lines show the behavior of the remainder component and the black lines the residuals once the remainder is modeled by such a process with $\phi_1 = 0.43$ and $\phi_2 = -0.47$, i.e. well inside the boundary region of the parameter space for stationarity. As it can be seen the AR(2) process captures most of the features in the structure of the remainder approximately up to lag 10. Nonetheless, some structure is still observed in lags 8, 13 and 16, which can be attributed to the rapid change in the slope of the underlying seasonal process, i.e. periodic singularities that are outside the scope of the method.

4.2 Weekly Stock Prices in Peru

We analyze the decomposition of a stock price in a typical illiquid financial market. The data records the closing price of a mining corporation devoted mainly to the extraction of cooper (with ticker MOROCOI1) at the end of each week between 01/05/1996 and 07/30/2010, or 761 observations. The prices are recorded in the original currency adjusted for rights, and when no trade takes place in a given week, a missing value is reported. The source of the data is Economatica.

The parameters for the model were chosen as in the previous example. Firstly, we consider 150 equidistant knots for both trend and seasonal components. We then relied on the inspection of the residuals in trend model (5). ACF and PACF analysis surprisingly suggest that the inclusion of frequency $\omega = 2\pi(760/126)$, and the use of an AR(1) process

to characterize the remainder are indeed appropriate. Furthermore, the application of the $\mathcal{R}(m_\tau)$ criterion presented in section 2 selects $m_\tau = 4$ as the optimal penalization choice in (4).

The first two panels in Figure 6 present the decomposition of the log prices during the analyzed period. The gray line in the upper-most panel depicts an oscillatory shape with an apparent dip at the beginning of 2003, and two utmost points at the beginning of years 1998 and 2008, all of which are synchronized with the dynamics of the international price of cooper in those years. These features are approximated by the fitted trend component (black line), and the addition of that and the seasonal part (dashed line), up to an apparent phase change in the seasonal component around year 2004, which coincides with the recovery in cooper prices due to a period of worldwide cooper shortage. The bottom panel of the same figure shows that AR(1) model imputed to the remainder is adequate. The gray lines indicate the behavior of the remainder before the AR(1) structure is included, and the black lines show the behavior of the residuals once the time series model is considered. The autocorrelation parameter in the latter was found to be $\phi = 0.2$.

An obvious argument against applying a decomposition algorithm to prices in financial markets is the efficient market hypothesis proposed by Fama (1970), which states that if indeed any structure existed in the market, traders would take bullish or bearish positions accordingly, and this action would make such structure vanish. Nonetheless, there is also empirical evidence suggesting that such hypothesis may be less reliable in developing countries (see e.g. Delgado and Humala, 1997; Kyaw, 2003; Worthington and Higgs, 2003). This has been argued to be the case, because the fluctuations in such markets show much more structure than those of developed countries. Some of the reasons for this feature are, for example, the fact that these markets deal with small transaction volumes; are mostly driven by institutional agents; and are typically illiquid.

4.3 Monthly Prices of Pork Meat in the EU

For this application we consider monthly data for the price of pork meat in the European Union between 08/1995 and 07/2012, that is 204 observations. The source of the data is the statistical office of the European Union (EUROSTAT), and records the average of grade E pig meat in cents per kilogram. To compute the average, the information is initially collected at a country level by the corresponding statistical authority, and is then reported to EUROSTAT where it is consolidated, its comparability is ensured and the weighted average (on production level per country) is calculated.

For the number of knots we consider $k_\tau = k_\alpha = 20$ only because of the relatively small number of observations, and the limited degrees of freedom vailable after the inclusion of six seasonal covariates (2 covariates per frequency). The other parameters are selected, as in the previous examples, following the steps presented in section 2. The analysis of the residuals of (5) given $m_\tau^* = 5$ reveals that the inclusion of frequencies $\omega_1 = 2\pi(204/12)$, $\omega_2 = 2\pi(204/6)$ and $\omega_3 = 2\pi(204/4)$, is adequate with no need for the use of an ARMA model for the remainder part. We do not find this result surprising considering that the data records monthly averages, and that for such resolution most of the short memory in remainder is expected to vanishe. The $\mathcal{R}(m_\tau)$ criterion selects a penalization parameter $m_\tau = 2$ for the trend component.

Once again, we present the decomposition of the time series under analysis in the first two panels of Figure 7. The data signal is presented as a gray line in the upper-most panel, and exhibits a great amount of structure. This is caused by the fact that it records monthly averages, and hence little variability is observed. As it can be seen, the estimated trend follows the structure of the data well, exhibiting a highly non-linear pattern which is possibly linked to the so-called *hog cycle* (see e.g. Ezequiel, 1938; Hayes and Schmitz, 1987), and other supply and demand anomalies affecting the market during the study period. As an example, it is interesting to consider the spiral of price increments at the

end of 1995 caused by the rise in pork meat demand consequence of the so-called *mad cow disease*; and the latter plumping of prices caused by the saturation of the market by producers eager to profit from the initial extraordinary benefits, which is considered the typical production dynamic driving the hog cycle's contraction phase.

With respect to the seasonal component, an apparent reduction in its amplitude along the analyzed years is reported in the middle panel of Figure 7. We do not find this observation surprising since there are arguments to believe that nowadays pork meat production has become less dependent on the yearly seasons mainly because of a contraction coming from the supply side of the market. Consider, for example, the effect of the globalization of feed markets that isolates the seasonal effect of the post harvest feeds, or the standardization of pork breeding practices in the industry under which pigs are now held in very large, isolated climate controlled units, making weather conditions almost negligible on the meat production side.

5 Conclusion

We presented a trend-seasonal decomposition method of time series based on penalized splines that considers varying seasonality and a non-trivial correlation structure for the remainder. The former is modelled by a varying coefficient model and the latter by an $ARMA(p, q)$ process. The method allows for an instant estimation of all parameters by means of the mixed model representation of penalized splines, which can be easily implemented by using the `R` package `nlme`. All together, the main advantages of our approach are the simultaneous modeling of all three components of a time series, the data-driven choice of smoothing and penalization parameters, as well as fast and simple numerical implementation. Real data examples and comparison to the STL method of Cleveland et al. (1990) confirmed practical relevance and effectiveness of the developed method.

Appendix: Technical details

Given the solution of (3), one can estimate the smooth part of the decomposition as $\hat{\boldsymbol{f}} = \boldsymbol{GY}$ with smoothing matrix $\boldsymbol{G} = \boldsymbol{C}(\boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}^t\boldsymbol{R}^{-1}$, where $\boldsymbol{C} = [\boldsymbol{C}_\tau, \boldsymbol{C}_\varsigma]$, and $\boldsymbol{D} = \text{blockdiag}\{\boldsymbol{D}_\tau, \boldsymbol{D}_\varsigma\}$ such that $\boldsymbol{C}_\tau = [\boldsymbol{X}_\tau, \boldsymbol{Z}_\tau]$, $\boldsymbol{C}_\varsigma = [\text{diag}\{\cos(\omega\boldsymbol{x})\}\boldsymbol{X}_\alpha, \text{diag}\{\cos(\omega\boldsymbol{x})\}\boldsymbol{Z}_\alpha]$; $\boldsymbol{D}_\tau = \sigma^2/\sigma^2_{u_\tau}\text{diag}\{\boldsymbol{0}_{m_\tau}, \boldsymbol{1}_{k_\tau+m_\tau}\}$ and $\boldsymbol{D}_\varsigma = \sigma^2/\sigma^2_{u_\alpha}\text{diag}\{\boldsymbol{0}_{m_\alpha}, \boldsymbol{1}_{k_\alpha+m_\alpha}\}$. The estimation of each component requires then a decomposition of $\boldsymbol{G}$, for which we can write: $\boldsymbol{G}_\tau = [\boldsymbol{C}_\tau, \boldsymbol{0}](\boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}^t\boldsymbol{R}^{-1}$ and $\boldsymbol{G}_\varsigma = [\boldsymbol{0}, \boldsymbol{C}_\varsigma](\boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{D})^{-1}\boldsymbol{C}^t\boldsymbol{R}^{-1}$ so that $\boldsymbol{G} = \boldsymbol{G}_\tau + \boldsymbol{G}_\varsigma$. Using formulae for the inverse of a partitioned matrix, and following the results by Aerts et al. (2002) we note that these smoothing matrices can also be written as:

$$\boldsymbol{G}_j = \boldsymbol{C}_j\{\boldsymbol{C}_j^t\boldsymbol{R}^{-1}(\boldsymbol{I} - \boldsymbol{S}_{-j})\boldsymbol{C}_j + \boldsymbol{D}_j\}^{-1}\boldsymbol{C}_j^t\boldsymbol{R}^{-1}(\boldsymbol{I} - \boldsymbol{S}_{-j}), j \in \{\tau, \varsigma\},$$

where the subscript $-j$ equals τ if j is set to ς and viceversa; and $\boldsymbol{S}_{-j} = \boldsymbol{C}_{-j}(\boldsymbol{C}_{-j}^t\boldsymbol{R}^{-1}\boldsymbol{C}_{-j} + \boldsymbol{D}_{-j})^{-1}\boldsymbol{C}_{-j}^t\boldsymbol{R}^{-1}$ is the smoothing matrix in a model with only component j.

References

Aerts, M., Claeskens, G., and Wand, M. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, 103:455–470.

Alexandrov, T., Bianconcini, S., Dagum, E., Maass, P., and McElroy, T. (2012). A review of some modern approaches to the problem of trend extraction. *Econometric Reviews*, 31(6):593–624.

Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(6):529–544.

Cleveland, R., Cleveland, W., McRae, J., and Terpenning, I. (1990). STL: A seasonal–trend decomposition procedure based on LOESS. *Journal of Official Statistics*, 6(1):3–73.

Cleveland, W. (1979). Robust locally regression and smoothing scatterplots. *Journal of*

the American Statistical Association, 74:829–836.

Dagum, E. (1978). Modelling, forecasting and seasonally adjusting economic time series with the X-11 ARIMA method. *Journal of the Royal Statistical Society. Series D*, 27(3):203–216.

Delgado, L. and Humala, A. (1997). El mercado bursátil peruano y la hipótesis del mercado eficiente. *Revista de Estudios Económicos*, 10.

European Union (2012). Statistical office. `http://epp.eurostat.ec.europa.eu`.

Ezequiel, M. (1938). The cobweb theorem. *Journal of Economics*, 52:255–280.

Fahrmeir, L. and Kneib, T. (2008). On the identification of trend and correlation in temporal and spatial regression. In Shalabh and Heumann, C., editors, *Recent Advances in Linear Models and Related Areas*, pages 1–27. Physica-Verlag.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2).

Findley, D., Monsell, B., Bell, W., Otto, M., and Chen, B. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Businness and Economic Statistics*, 16(2):127–177.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, USA.

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55:756–796.

Hayes, D. and Schmitz, A. (1987). Hog cycles and countercyclical production response. *American Journal of Agricultural Economics*, 69:762–770.

Henderson, R. (1916). Note on graduation by adjusted average. *Transactions of the American Society of Actuaries*, 17:43–48.

Hodrick, R. and Prescott, E. (1997). Postwar u.s. business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, 29(1):1–16.

Kauermann, G., Krivobokova, T., and Semmler, W. (2011). Filtering time series with

penalized splines. *Studies in Nonlinear Dynamics & Econometrics*, 15(2).

Krivobokova, T. (2012). Smoothing parameter selection in two frameworks for penalized splines. *Journal of the Royal Statistical Society, Series B.*, to appear.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480):1328–1337.

Kyaw, N. (2003). Persistence characteristics of latin american financial markets. Technical report. Working Paper, Kent State University.

Maravall, A. and Caporello, G. (2004). Program TSW: Revised reference manual. Technical Report, Research Dep., Bank of Spain. http://www.bde.es.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16:134–153.

Pollock, D. (2006). Econometric methods of signal extraction. *Computational Statistics and Data Analysis*, 50(9):2268–2292.

Pollock, D. and Cascio, I. L. (2007). Non-dyadic wavelet analysis. In Kontogoghiorges, E. and Gatu, C., editors, *Optimisation, Econometric and Financial Analysis*, pages 167–203. Springer-Verlag.

Ramsay, J. and Lampart, C. (1998). The decomposition of economic relationships by time scale using wavelets: Expenditure and income. *Studies in Nonlinear Dynamics & Econometrics*, 3(1).

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*. Cambridge University Press, New York.

Wand, M. (2003). Smoothing and mixed models. *Computational statistics*, 18(2):223–249.

Wiesenfarth, M., Krivobokova, T., Klasen, T., and Sperlich, S. (2012). Direct simultaneous inference in additive models and its application to model undernutrition. *Journal of the American Statistical Association*, to appear.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.

Worthington, A. and Higgs, H. (2003). Tests of random walk and market efficiency in latin american stock markets: An empirical note. Technical report. Working paper, Queenslnad University of Technology.

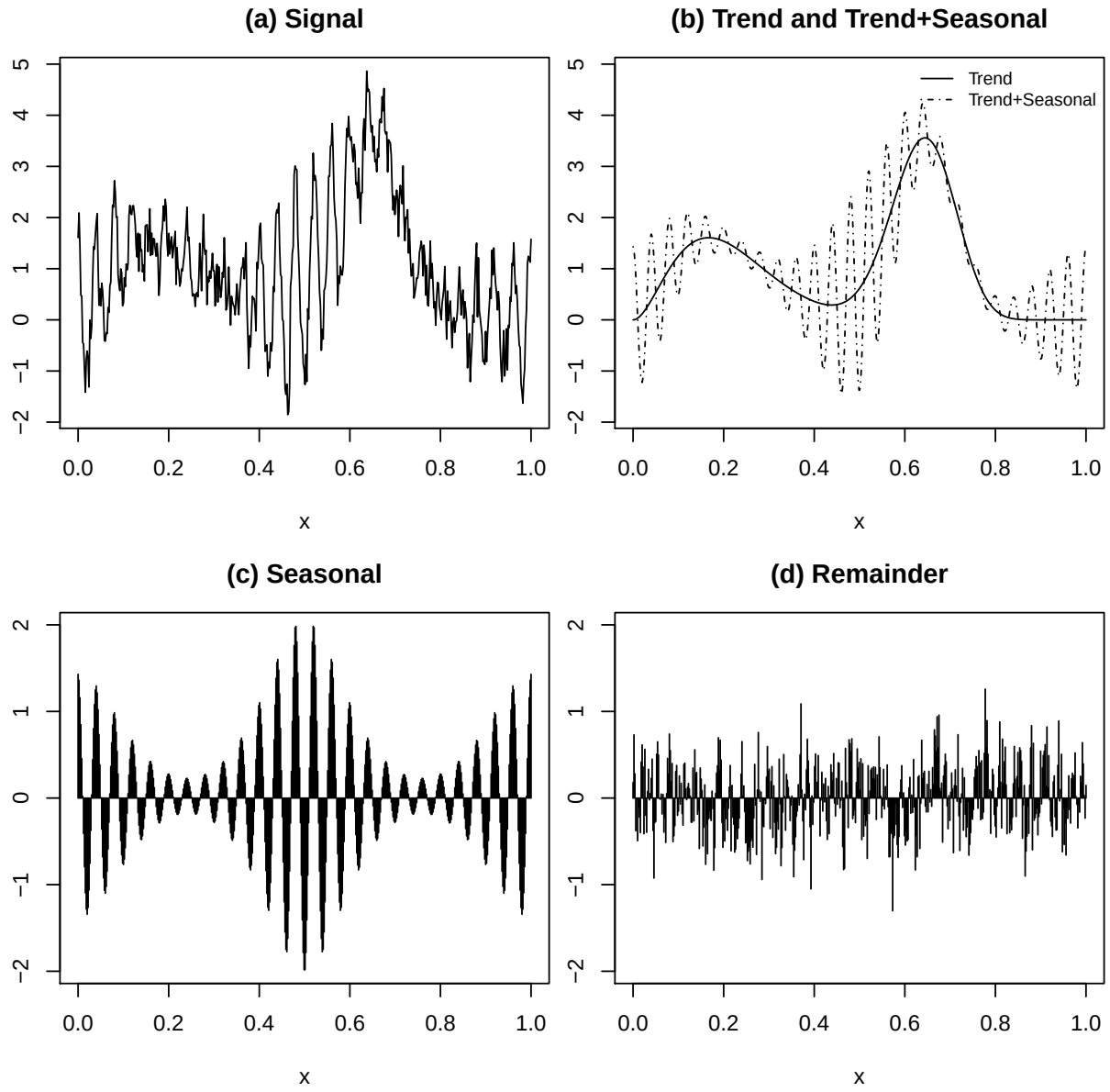

Figure 1: Simulation Example. (a) time series, (b) trend component, (c) seasonal component and (d) remainder component.

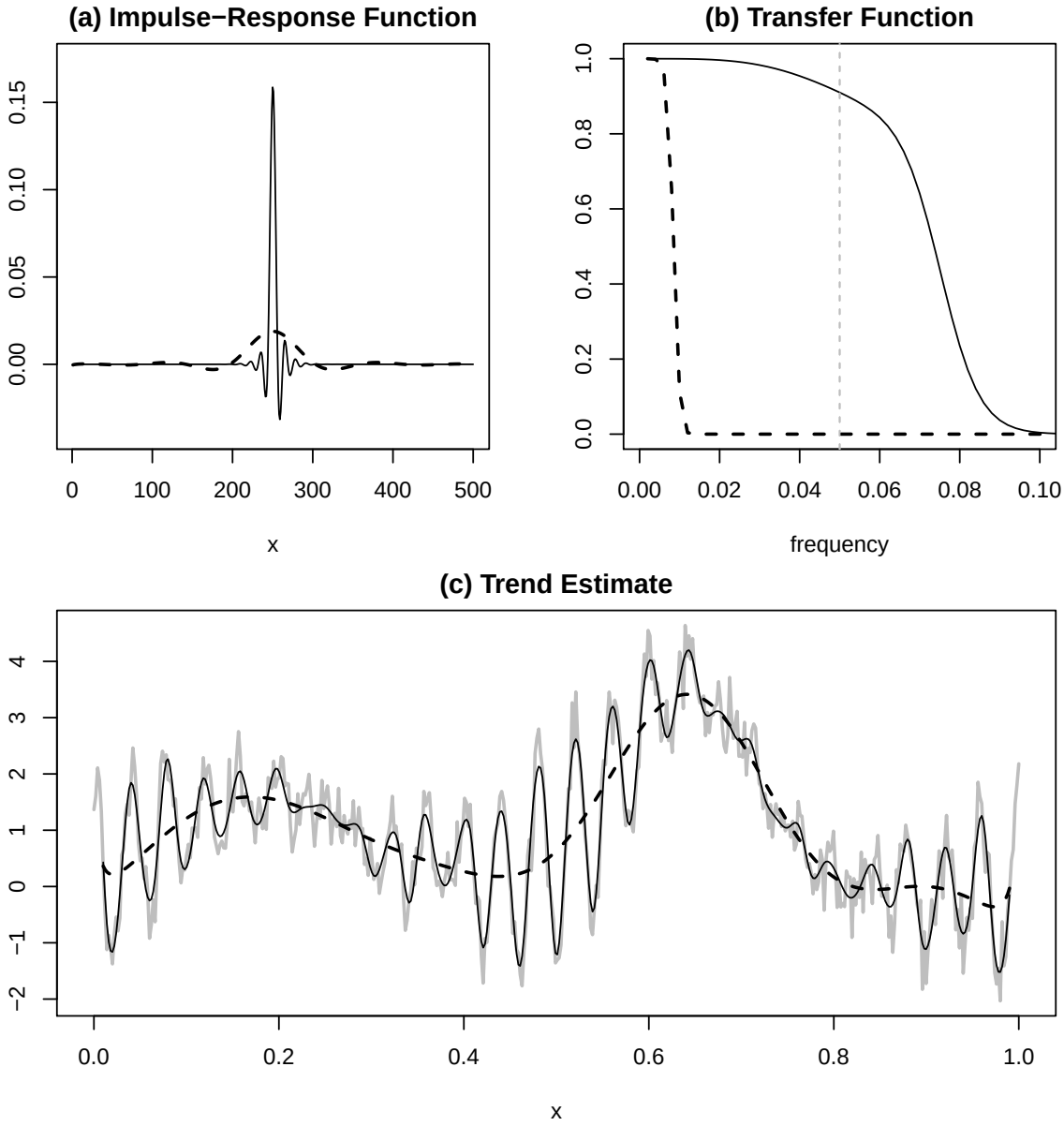

Figure 2: (a) Impulse-Response Function; (b) Transfer Function; (c) Trend estimate. In all plots the continuous and dashed lines represent the cases when $m_\tau = 2$ and $m_\tau = 5$ respectively. In (c) the simulated data is added as a gray line.

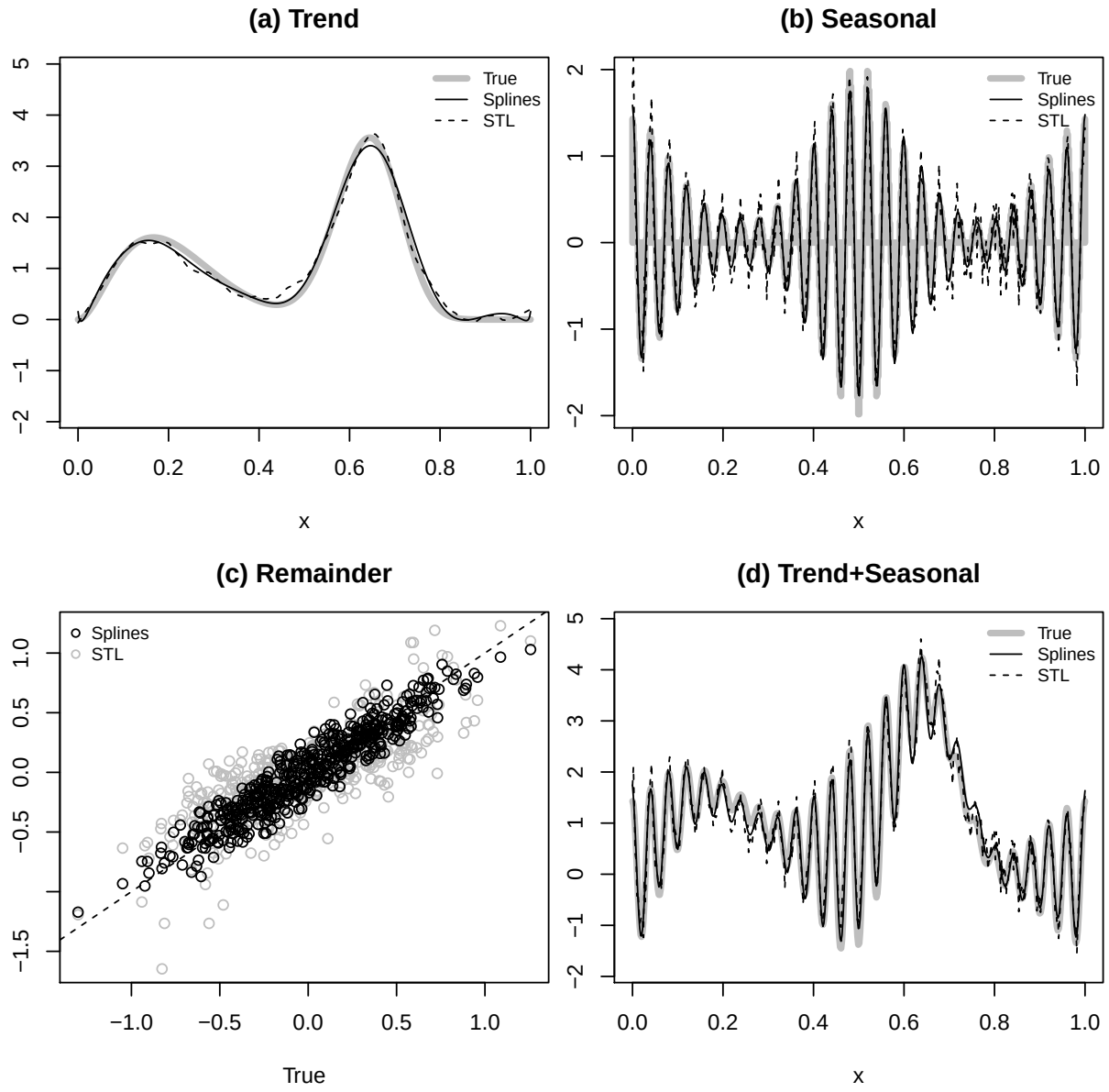

Figure 3: Comparison between STL and penalized splines method. For panels (a), (b) and (d) the dashed and the continuous lines show the results for the STL and spline methods respectively: (a) trend component comparison, (b) seasonal component comparison, and (d) added trend and seasonal estimation comparison. Panel (c) shows the scatter plots of the resulting remainder components from the simulations and the real ones for the STL (gray circles) and penalized splines methods (black circles).

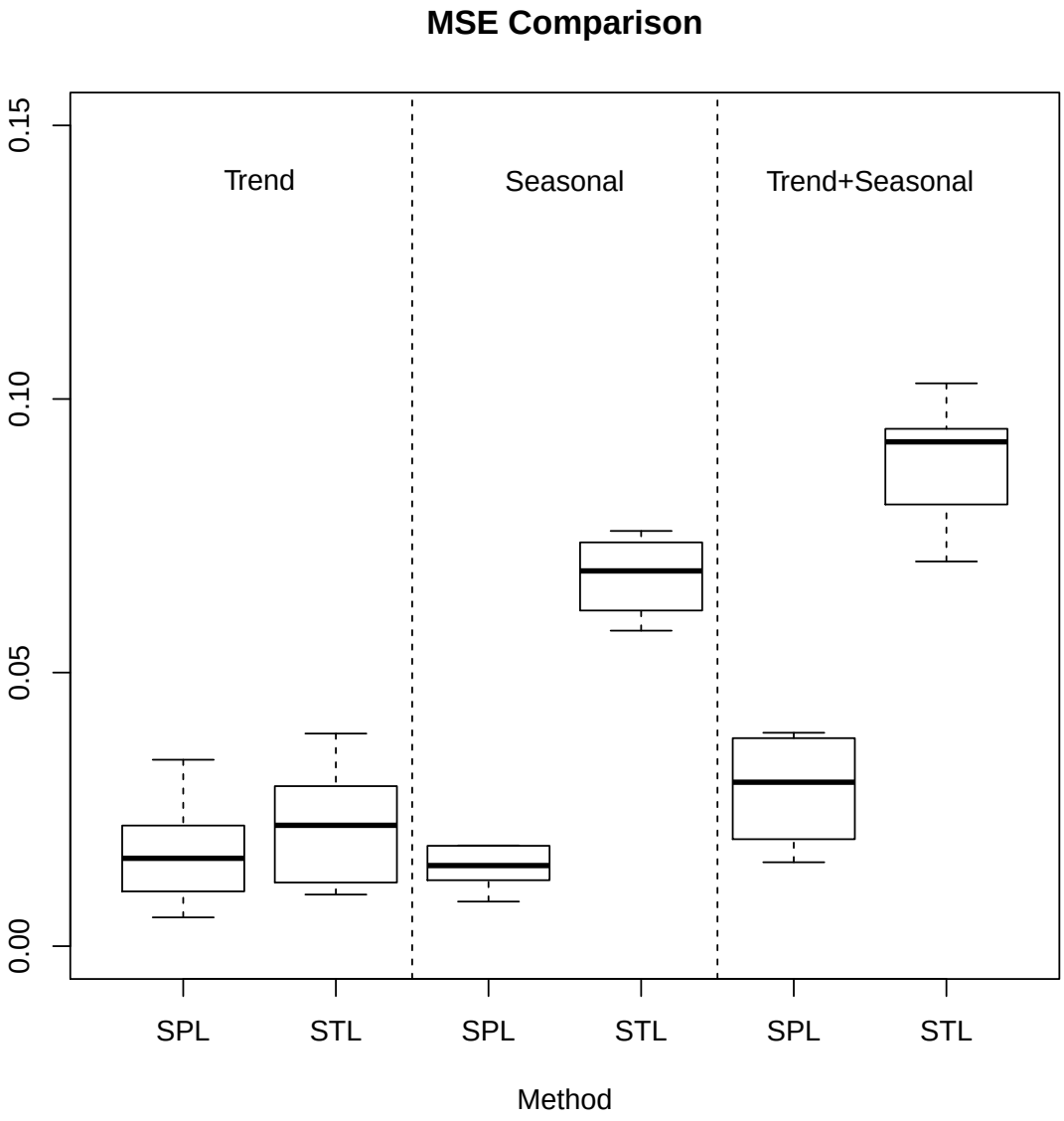

Figure 4: MSE component-wise comparison between the spline method (SPL in the figure) and the STL procedure in the Monte Carlo experiment.

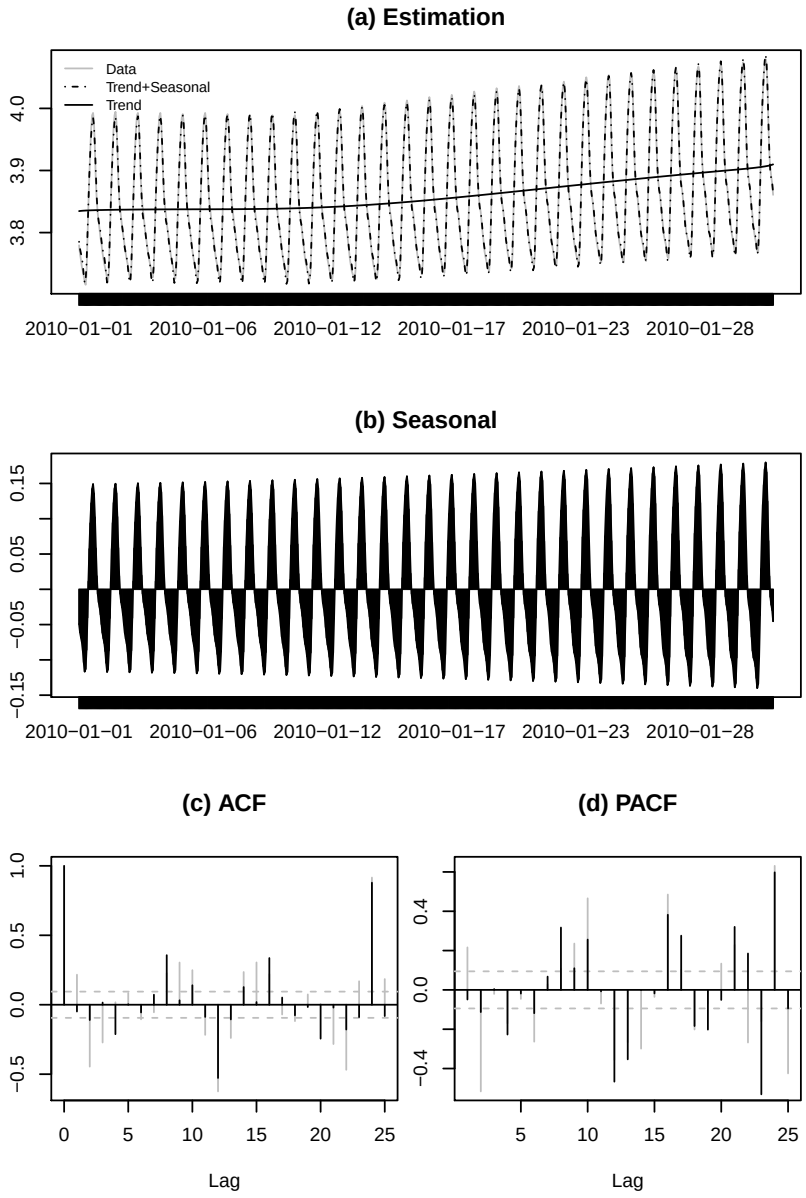

Figure 5: Decomposition Example for Hourly Temperatures in California. (a) Raw data (gray line) and estimators for the trend (black continuous line), and the added trend and seasonal estimations (black dashed line); (b) estimated seasonal component; and (c) ACF and PACF for the model with (black lines) and without (gray lines) an explicit characterization of the remainder.

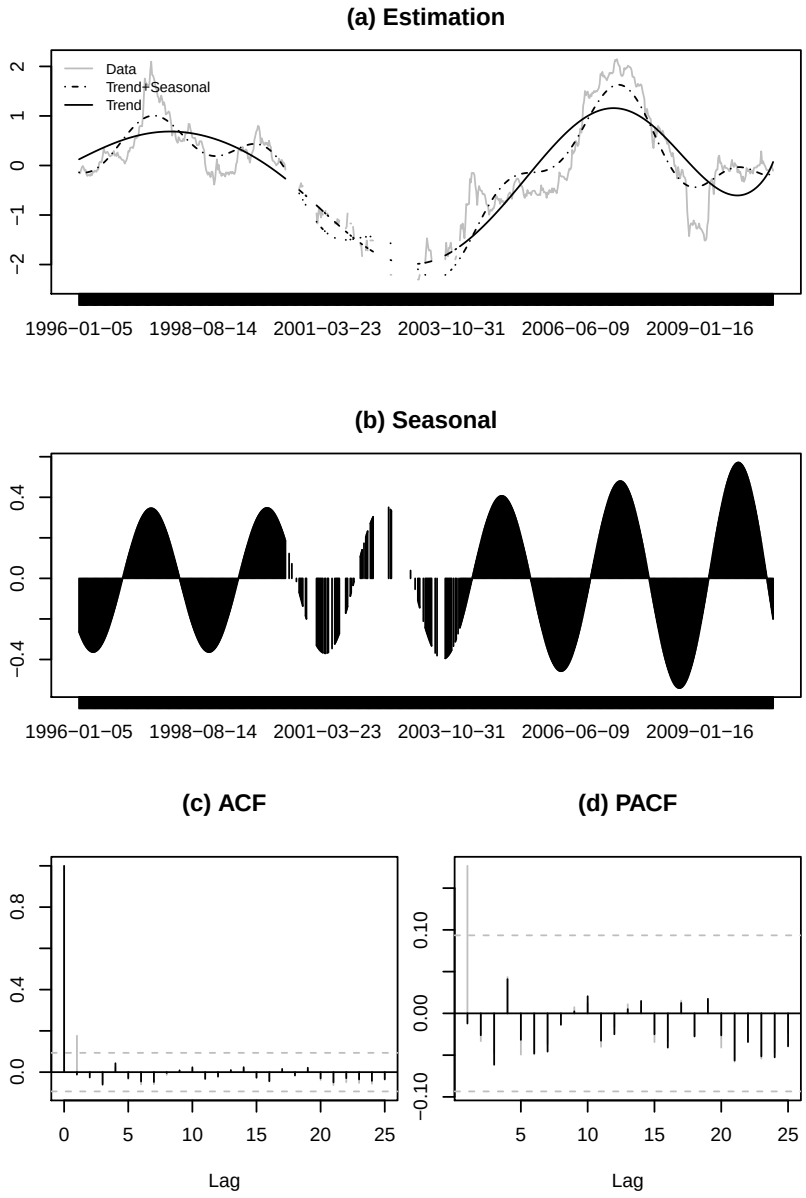

Figure 6: Decomposition Example for Weekly Stock Prices in Peru. (a) Raw data (gray line) and estimators for the trend (black continuous line), and the added trend and seasonal estimations (black dashed line); (b) estimated seasonal component; and (c) ACF and PACF for the model with (black lines) and without (gray lines) an explicit characterization of the remainder.

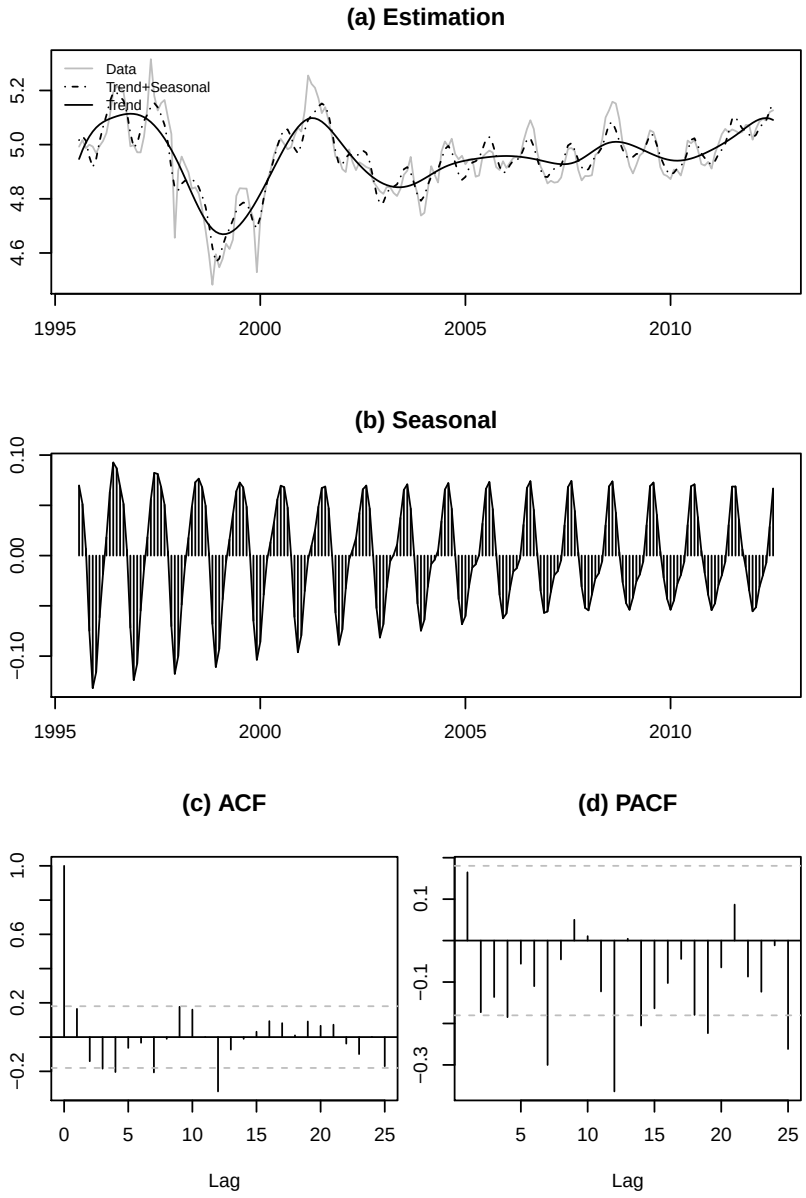

Figure 7: Decomposition Example for Monthly Pork Meat Prices in the EU. (a) Raw data (gray line) and estimators for the trend (black continuous line), and the added trend and seasonal estimations (black dashed line); (b) estimated seasonal component; and (c) ACF and PACF for the model with (black lines) and without (gray lines) an explicit characterization of the remainder.