On Modeling Heterogeneity in Linear Models using Trend Polynomials

Michael Michaelides[†] [*]
[†]Allegheny College,
Department of Economics,
Quigley Hall, Box 20,
Meadville, PA 16335, USA.
E-mail address: mmichaelides@allegheny.edu

Aris Spanos[‡]
[‡]Virginia Tech,
Department of Economics,
Pamplin Hall, Mail Code 0316,
Blacksburg, VA 24061, USA.
E-mail address: aris@vt.edu

Published in *Economic Modelling,* Volume 85, February 2020, 74-86

Submitted: 15 October 2018

Article Accepted for publication: 10 May 2019

Abstract

The primary aim of the paper is to consider the problems and issues raised when the data exhibit time heterogeneity in the context of linear models. Ignoring time heterogeneity will undermine the reliability of inference and will give rise to untrustworthy evidence. Accounting for it using trend polynomials, however, is non-trivial because it raises several modeling issues. First, when the degree of the polynomial is greater than 4, or so, one needs to deal with the near-multicollinearity problem that arises. The second issue pertains to the type of polynomial that will adequately account for the time heterogeneity. Third, when the trend polynomials are treated as additional regressors, they will give rise to highly misleading statistical results. The paper investigates how different types of polynomials could deal with the near-multicollinearity and the modeling issues they raise, and makes recommendations to practitioners.

Key words: linear model; t-heterogeneity; near-collinearity; trend polynomial; orthogonal polynomial; orthonormal polynomial.

JEL: C18, C22, C51, C52, C58

[*]Correspondence author: Michael Michaelides, 520 N. Main Street, Box 20, Department of Economics, Allegheny College, Meadville, PA 16335, USA; phone: +1 (814) 332-3346; e-mail address: mmichaelides@allegheny.edu.

1 Introduction

Beginning in the late 19th century, the modeling of economic time series was primarily based on the ad hoc decomposition (Morgan, 1990):

$$y_t = \text{trend} + \text{seasonal} + \text{cycles} + \text{noise}, \; t{=}1, 2, ..., n, ... \tag{1}$$

This was motivated by the fact that even a casual inspection of a *t-plot* (time plot) of most economic time series reveals three distinct regularity patterns: a trend, different cycles, and a certain degree of jaggedness around the trend and cycles that was viewed as random noise. When the cycles seem very regular, they are often interpreted as seasonality, but when they are irregular are viewed as business cycles.

The introduction of correlation and regression analysis by Galton in the 1880s (Stigler, 1986) provided the key tools for modeling time series data using (1). It soon became apparent, however, that correlation and regression analysis could give rise to 'dubious' results: apparent statistical associations that after closer scrutiny of the data are rendered 'spurious'; see Hooker (1901) and Yule (1903). This led practitioners to use several ad hoc methods to remove the trend component with a view to address the problem of statistically spurious results. The most popular was to first remove the trend in some way, usually using deterministic trend polynomials up to degree 3 (Moore, 1914), and then apply correlation/regression to the detrended series. An alternative procedure, suggested by Hooker (1905), was to first difference the data series and then apply correlation/regression using the first differences. It was not, however, obvious to which extend such methods addressed the spuriousness of the ensuing statistical results. Indeed, the problem of spurious correlation and regression results, initially raised by Yule (1926), resurfaced in Granger and Newbold (1974) using simulated data based on unit root processes. Despite the elucidating technical account proposed by Phillips (1986) to explain their simulation results, the real problem of addressing the spuriousness remained unresolved. To this day, this remains a problem that has not been adequately addressed, notwithstanding the extraordinary developments in statistics and econometrics since then.

In empirical modeling the statistical systematic information comes in the form of chance regularity patterns exhibited by the data. When evaluated in terms of learning from data about phenomena of interest, the success of such modeling depends crucially on how adequately these patterns are accounted for by the prespecified statistical model that comprises probabilistic assumptions from three broad categories: *Distribution*, *Dependence*, and *Heterogeneity*; see Spanos (2019). The heterogeneity category has been the least developed in the sense of devising concepts to account for different patterns of heterogeneity in economic data; see Phillips (2005). There are several reasons for this neglect, including the following.

First, over the last quarter century the focus on unit roots (Choi, 2015) has created the impression that the scope of unit root heterogeneity (stochastic trends) is broad enough to account for the majority of economic time series. In fact unit root type heterogeneity is of rather limited scope. Focusing on the first two moments for simplicity, a stochastic process $\{Z_t,\ t\in\mathbb{N}:=(1,2,...,n,...)\}$ is time heterogeneous when:

$$E(Z_t)=\mu(t),\ Cov(Z_t,Z_s)=E[Z_t-\mu(t)][Z_s-\mu(s)]=v(t,s),\ \text{for all } t,s\in\mathbb{N}, \tag{2}$$

where $(\mu(t),v(t,s),\ t,s\in\mathbb{N})$ are arbitrary functions of t and s. The unit root non-stationarity that has dominated time series modeling since the early 1990s is an extremely special case of (2) where the arbitrary functions $\mu(t)$ and $v(t,s)$ take very simple forms:

$$\mu(t) = \mu{\cdot}t \text{ and } v(t,s)=\sigma(0)\min(t,s),\ \text{for all } t,s\in\mathbb{N}. \tag{3}$$

Due to its highly restrictive form, one would expect only a small fraction of *t-heterogeneity* (time heterogeneity) in economic time series data to be adequately modeled using stochastic trends stemming from unit root models, the simplest of which is:

$$y_t = y_{t-1}+\varepsilon_t \rightarrow y_t = y_0 + \overbrace{\sum\nolimits_{j=1}^{t}\varepsilon_j}^{\text{stochastic trend}},\ \varepsilon_t\backsim\mathsf{NIID}(0,\sigma^2),\ t\in\mathbb{N}. \tag{4}$$

Second, accounting for t-heterogeneity using generic trend polynomials in linear regression and related models is hampered by practical difficulties when the degree of the polynomial (p) is greater than 4 or so. It is well known that such polynomials are likely to give rise to *near-collinearity* (near-multicollinearity) problems; see Greene (2011, p. 129). Complicated forms of t-heterogeneity can easily arise in cases where several data series with heterogeneity are aggregated; see section 5 for an empirical example. In practice, it is not always easy to detect the presence of such t-heterogeneity by just glancing at the t-plots of the data. Moreover, it is uncertain what type of trend polynomials is likely to adequately account for such t-heterogeneity.

In general, accounting for t-heterogeneity exhibited by observed data is essential in empirical modeling because the presence of a neglected trend in one's data can result in inconsistent estimators for the parameters of interest, as well as sizeable discrepancies between the nominal (assumed) error probabilities (type I, II, power and coverage) and the actual ones. The primary aim of this paper is to articulate the potential consequences of neglecting the presence of t-heterogeneity in the data, and discuss the practical difficulties arising from accounting for this misspecification in the context of linear models. The discussion focuses on exploring different forms of trend polynomials whose degree could be potentially extended to $p>4$ and propose ways to address the potential near-collinearity problems that are likely to arise. Even though the focus of this paper is on t-heterogeneity, the discussion applies equally to cross-section data whose level of measurement is at least of ordinal scale.

To motivate the discussion that follows, section 2 brings out the serious consequences on the reliability of inference by not accounting for the mean t-heterogeneity in one's data. Section 3 discusses the problem of near-collinearity in the context of a linear model stemming from the ill-conditioning of the regressor matrix. Section 4 evaluates a number of different ways one could deal with the near-collinearity problem. These include ordinary trend polynomials, the most widely used continuous and discrete orthogonal trend polynomials, orthonormal trend polynomials, and scaled trend polynomials. It is shown that conventional approaches to the modeling of t-heterogeneity do not adequately address the near-collinearity problem. Section 5 illustrates how one can model complicated forms of t-heterogeneity using an empirical example from asset pricing theory. Finally, section 6 concludes with recommendations to the practitioner.

2 Statistical misspecification: mean heterogeneity

2.1 The Linear Regression (LR) model: specification

In an attempt to bring out the serious consequences for the reliability of inference stemming from ignoring the mean t-heterogeneity ($\mu(t)$: a trending mean) in one's data, let us focus the discussion on the Linear Regression (LR) model:

$$\begin{gathered} y_t{=}\delta_0{+}\delta_1 t{+}\delta_2 t^2{+}\beta_1 x_{1t}{+}\beta_2 x_{2t}{+}u_t, \\ (u_t|X_{1t}{=}x_{1t}, X_{2t}{=}x_{2t}) \backsim \mathsf{NIID}(0,\sigma^2),\ \ t{\in}\mathbb{N}, \end{gathered} \tag{5}$$

where 'NIID' stands for 'Normal, Independent and Identically Distributed'. The LR model is used because it can be easily extended to the autoregressive of order p (AR(p)) model, which provides the cornerstone for time series modeling.

The error assumptions imply a particular statistical parameterization for the unknown parameters $\boldsymbol{\theta}{:=}(\delta_0,\delta_1,\delta_2,\beta_1,\beta_2,\sigma^2)$ in terms of the moments of the observable process $\{\mathbf{Z}_t{:=}(y_t,X_{1t},X_{2t}),\ t{\in}\mathbb{N}\}$ underlying data $\mathbf{Z}_0$; see Spanos and McGuirk (2002). One can derive the parameterization directly using the joint distribution of the observable random variables involved:

$$\begin{pmatrix} y_t \\ X_{1t} \\ X_{2t} \end{pmatrix} \backsim \mathsf{NIID}\left(\begin{pmatrix} c_{01}{+}c_{11}t{+}c_{21}t^2 \\ c_{02}{+}c_{12}t{+}c_{22}t^2 \\ c_{03}{+}c_{13}t{+}c_{23}t^2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}\right). \tag{6}$$

In this case, the regression and skedastic functions take the form (table 1):

$$E(y_t|\mathbf{X}_t{=}\mathbf{x}_t){=}\beta_0(t){+}\beta_1 x_{1t}{+}\beta_2 x_{2t},\ \ Var(y_t|\mathbf{X}_t{=}\mathbf{x}_t){=}\sigma^2,$$

where $\beta_0(t){=}\delta_0{+}\delta_1 t{+}\delta_2 t^2$, $\mathbf{X}_t{:=}(X_{1t},X_{2t})^\top$, and the parameterizations of $\boldsymbol{\theta}$ are (table 1):

$$\beta_0(t){=}E(y_t){-}\beta_1 E(X_{1t}){-}\beta_2 E(X_{2t}),\ \ \beta_1{=}\tfrac{(\sigma_{12}\sigma_{33}-\sigma_{13}\sigma_{23})}{(\sigma_{22}\sigma_{33}-\sigma_{23}^2)},\ \ \beta_2{=}\tfrac{(\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23})}{(\sigma_{22}\sigma_{33}-\sigma_{23}^2)}, \tag{7}$$

$$\sigma^2{=}\sigma_{11}{-}\sigma_{12}\beta_1{-}\sigma_{13}\beta_2, \quad \mathcal{R}^2{=}1{-}\frac{\sigma^2}{\sigma_{11}}. \tag{8}$$

Table 1 specifies the LR model in terms of the Statistical Generating Mechanism (GM) and assumptions [1]-[5] that constitute a complete, internally consistent and testable set of assumptions in terms of the observable process $\{(y_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$ underlying the data $\mathbf{Z}_0{:=}\{(y_t,\mathbf{x}_t),\ t{=}1,2,...,n\}$. This provides a purely probabilistic construal of a statistical model, viewed as a particular parameterization of the process $\{(y_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$. Intuitively, the statistical model comprises the totality of probabilistic assumptions one imposes on the process $\{(y_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$ with a view to render data $\mathbf{Z}_0$ a 'typical' realization thereof. The 'typicality' is testable using comprehensive misspecification testing. It is important to note that the specification of the LR model in table 1 can be related directly to the traditional specification in terms of the error process $\{(u_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$; see Spanos (2019) for further discussion.

Table 1: Linear Regression (LR) Model

Statistical GM:		$y_t{=}\beta_0{+}\boldsymbol{\beta}_1^\top\mathbf{x}_t{+}u_t,\ t{\in}\mathbb{N}.$	
[1]	Normality:	$(y_t|\mathbf{X}_t{=}\mathbf{x}_t)\backsim \mathsf{N}(.,.),$	$t{\in}\mathbb{N}.$
[2]	Linearity:	$E(y_t|\mathbf{X}_t{=}\mathbf{x}_t){=}\beta_0{+}\boldsymbol{\beta}_1^\top\mathbf{x}_t,$	
[3]	Homoskedasticity:	$Var(y_t|\mathbf{X}_t{=}\mathbf{x}_t){=}\sigma^2,$	
[4]	Independence:	$\{(y_t|\mathbf{X}_t{=}\mathbf{x}_t),\ t{\in}\mathbb{N}\}$ indep. process,	
[5]	Parameter constancy:	$(\beta_0,\boldsymbol{\beta}_1,\sigma^2)$ are *not* changing with t,	
	$\beta_0{=}E(y_t){-}\boldsymbol{\beta}_1^\top E(\mathbf{X}_t),\ \boldsymbol{\beta}_1{=}[Cov(\mathbf{X}_t)]^{-1}Cov(\mathbf{X}_t,y_t),$		
	$\sigma^2{=}Var(y_t){-}Cov(\mathbf{X}_t,y_t)^\top[Cov(\mathbf{X}_t)]^{-1}Cov(\mathbf{X}_t,y_t).$		

2.2 Mean heterogeneity and the reliability of inference

In this section, we use a Monte Carlo simulation to trace the implications of ignoring the mean t-heterogeneity on the reliability of any inference based on the least-squares (OLS) estimators of $\boldsymbol{\theta}{:=}(\delta_0,\delta_1,\delta_2,\beta_1,\beta_2,\sigma^2)$. Using the simulation design based on the following joint distribution:

$$\begin{pmatrix} y_t \\ x_{1t} \\ x_{2t} \end{pmatrix} \backsim \mathsf{NIID}\left(\begin{pmatrix} 2.5{+}.4t{+}.12t^2 \\ 2{+}.2t{+}.22t^2 \\ 1.5{+}.15t{+}.25t^2 \end{pmatrix}, \begin{pmatrix} 1.5 & .75 & .65 \\ .75 & 1.1 & .2 \\ .65 & .2 & 1 \end{pmatrix}\right),$$

the statistical parameterizations in (7)-(8) give rise to the following:

$$(\beta_1,\beta_2)=(.584,.533),\ \beta_0(t){=}.5325{+}.203t{-}.142t^2,\ \sigma^2{=}.714,\ \mathcal{R}^2{=}.524,$$

$$y_t{=}.5325{+}.203t{-}.142t^2{+}.584x_{1t}{+}.533x_{2t}{+}\sqrt{.714}\epsilon_t,\ t{=}1,2,...,n,$$

where $\epsilon_t{\backsim}\mathsf{NIID}(0,1)$ denotes pseudo-random numbers.

The results in table 2 refer to two scenarios: (a) statistically adequate case where the results are based on estimating the correct LR model in (5), and (b) a statistically misspecified case where the modeler ignored the presence of the trends and estimated $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$. The table summarizes the simulation results in terms of the mean and standard error (SE) of the sampling distributions of the OLS estimators and the t-tests, where the star (*) indicates the 'true' value.

Table 2: Simulation results for the Linear Regression (LR) model

Replications: N=10000	(a) Adequate model $y_t=\delta_0+\delta_1 t+\delta_2 t^2+\beta_1 x_{1t}+\beta_2 x_{2t}+u_t$				(b) Misspecified model $y_t=\beta_0+\beta_1 x_{1t}+\beta_2 x_{2t}+\varepsilon_t$			
Estimators	n=50		n=100		n=50		n=100	
of Parameters	Mean	SE	Mean	SE	Mean	SE	Mean	SE
$\hat{\beta}_0$ $[\beta_0^*=.533]$	0.523	0.472	0.528	0.325	4.204	0.382	6.117	0.481
$\widehat{\delta}_1$ $[\delta_1^*=.203]$	0.203	0.044	0.203	0.023	-	-	-	-
$\widehat{\delta}_2$ $[\delta_2^*=-.142]$	-0.142	0.038	-0.142	0.026	-	-	-	-
$\hat{\beta}_1$ $[\beta_1^*=.584]$	0.587	0.122	0.586	0.084	0.508	0.166	0.943	0.173
$\hat{\beta}_2$ $[\beta_2^*=.533]$	0.531	0.128	0.532	0.088	0.054	0.147	-0.340	0.152
$\hat{\sigma}^2$ $[\sigma_*^2=.714]$	0.715	0.151	0.714	0.106	2.066	0.362	5.100	0.513
R^2 $[R_*^2=.524]$	0.554	0.095	0.538	0.068	1.000	0.000	1.000	0.000
t-statistics	Mean	α=.05	Mean	α=.05	Mean	α=.05	Mean	α=.05
$\tau_{\beta_0}=\frac{(\hat{\beta}_0-\beta_0^*)}{\hat{\sigma}_{\beta_0}}$	-0.017	0.049	-0.013	0.050	9.758	1.000	11.679	1.000
$\tau_{\delta_1}=\frac{(\widehat{\delta}_1-\delta_1^*)}{\hat{\sigma}_{\delta_1}}$	0.012	0.051	-0.049	0.050	-	-	-	-
$\tau_{\delta_2}=\frac{(\widehat{\delta}_2-\delta_2^*)}{\hat{\sigma}_{\delta_2}}$	0.009	0.049	-0.009	0.051	-	-	-	-
$\tau_{\beta_1}=\frac{(\hat{\beta}_1-\beta_1^*)}{\hat{\sigma}_{\beta_1}}$	0.015	0.051	0.000	0.050	-0.526	0.074	2.189	0.575
$\tau_{\beta_2}=\frac{(\hat{\beta}_2-\beta_2^*)}{\hat{\sigma}_{\beta_2}}$	-0.017	0.051	0.014	0.049	-3.323	0.879	-5.802	1.000

(a) **Statistically adequate model**. (i) The point estimates are *highly accurate* because they come from *consistent* estimators of the corresponding parameters,

(ii) the empirical type I error probabilities associated to the t-tests are very close to the nominal (α=.05) even for a sample size n=50, and

(iii) the accuracy of both point estimators and the relevant error probabilities improves as n increases to n=100.

(b) **Statistically misspecified model**. (i) The point estimates are *highly inaccurate* because they are based on *inconsistent* estimators of the corresponding parameters. The sampling distribution of $\widehat{\beta}_0$ is shifted to the right and the displacement as well as the variance increase with n. The sampling distribution of $\hat{\beta}_1$ is shifted to the left of the true value for n=50, and then shifted to the right for n=100, with its variance increasing. The sampling distribution of $\hat{\beta}_2$ is shifted to the left of the true value for both sample sizes, but its sign changes from positive when n=50 to negative when n=100. The sampling distributions of $\left(\hat{\sigma}^2, R^2\right)$ are shifted to the right with the displacement as well as the variance increasing with n.

(ii) There are substantial discrepancies between the nominal type I error probability of .05 and the actual ones.

(iii) As n increases the inaccuracy of the estimates increases and the empirical type I error probabilities approach one.

Looking at the above simulation results, it is clear that ignoring the mean t-heterogeneity (scenario b) has devastating effects on the reliability of estimation and the testing results associated with the LR model. The problem is that when the data exhibit mean t-heterogeneity that is not accounted for, all the sample second and higher central moments will be inconsistent estimators of the corresponding distribution moments, leading to inconsistent estimators for the regression coefficients, as well as sizeable discrepancies between the nominal and actual error probabilities. Indeed, as the sample size increases the effects of t-heterogeneity on the reliability of inference become more and more pernicious. It is worth mentioning that the simulation results would not be different for simulation designs that involve linear, cubic, or higher degree trend polynomials.

One might object to the above simulation exercise on the grounds that it will be very easy to detect the trends by just glancing at the t-plots of the data. In practice, however, t-heterogeneity can easily arise in more subtle ways that cannot be easily detected by eyeballing. Indeed, one of the objectives of this paper is to bring out the fact that the presence or absence of t-heterogeneity is not just a matter of eyeballing the data t-plots, but it has to be decided on the basis of formal misspecification testing, by probing whether the residuals contain any form of t-heterogeneity. For instance, by eyeballing the data on aggregate stock portfolio returns in figures 1-2 and asset returns in figure 3 (section 5) one will be hard pressed to detect any form of mean trending. A more formal misspecification testing, however, reveals the presence of complicated high order t-heterogeneity that will invalidate any inferences based on any statistical models that ignore it.

3 Linear Models and trend polynomials

3.1 Near-collinearity in numerical analysis

From a numerical analysis perspective, the real issue associated with near-collinearity concerns the potential instabilities (wobbliness) of the numerical values of $\mathbf{b}=\mathbf{Z}^{-1}\mathbf{y}$ to small changes in the data $(\mathbf{Z},\mathbf{y})$, when solving the linear system $\mathbf{Zb}=\mathbf{y}$ for $\mathbf{b}$. In the context of the LR model, this problem arises when solving the first order conditions to estimate $\boldsymbol{\beta}^{\top}:=(\beta_0,\boldsymbol{\beta}_1^{\top})$:

$$\mathbf{X}^{\intercal}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0} \rightarrow \ \mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^{\intercal}\mathbf{y} \rightarrow \widehat{\boldsymbol{\beta}}=\left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\mathbf{X}^{\intercal}\mathbf{y}. \qquad (9)$$

From a numerical perspective, the potential instability of the solution $\widehat{\boldsymbol{\beta}}$ stems from the ill-conditioning of $(\mathbf{X}^{\intercal}\mathbf{X})$ in (9); see Gautschi and Inglese (1988), Gautschi (1990), and Tyrtyshinikov (1994).

A widely used measure to quantify the extend of ill-conditioning of any $(n \times p)$ matrix $\mathbf{X}$ is the norm condition number:

$$\kappa(\mathbf{X}) = \|\mathbf{X}\| \cdot \|\mathbf{X}^{-1}\|, \tag{10}$$

where $\|\cdot\|$ denotes a matrix norm. The most widely used norms for a matrix $\mathbf{X}$ are:

$$\text{Euclidean:} \quad \|\mathbf{X}\|_2 = \sup_{\mathbf{b}\neq\mathbf{0}} \tfrac{\|\mathbf{Xb}\|_2}{\|\mathbf{b}\|_2} = \sqrt{\lambda_{\max}},$$

$$\text{Frobenius:} \quad \|\mathbf{X}\|_F = \sqrt{\textstyle\sum_{i=1}^{n}\sum_{j=1}^{p}|x_{ij}|^2} = \sqrt{\text{trace}(\mathbf{X}^\top\mathbf{X})} = \sqrt{\textstyle\sum_{i=1}^{p}\lambda_i^2},$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, $(\lambda_1 = \lambda_{\max}, \lambda_p = \lambda_{\min})$ are the eigenvalues of the positive-semidefinite matrix $(\mathbf{X}^\top\mathbf{X})$. These two norms are related via (Golub and Van Loan, 2013):

$$\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F \leq \sqrt{p}\,\|\mathbf{X}\|_2 .$$

The eigenvalues stem from the Singular Value Decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{U}^\top\mathbf{X}\mathbf{V} = \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p),$$

where $\mathbf{U}$ and $\mathbf{V}$ are $(n \times n)$ and $(p \times p)$ matrices, respectively.

It is important to note that the eigenvalues λ_i, $i=1,2,..,p$, of any data matrix $\mathbf{X}$, rank$(\mathbf{X})=p$, are directly related to those of $(\mathbf{X}^\intercal\mathbf{X})$, which are λ_i^2, $i=1,2,..,p$. Hence, the Euclidean and Frobenius condition numbers for $(\mathbf{X}^\intercal\mathbf{X})$ are:

$$\begin{aligned} &\kappa_2(\mathbf{X}^\intercal\mathbf{X}) = \|(\mathbf{X}^\intercal\mathbf{X})\|_2 \cdot \|(\mathbf{X}^\intercal\mathbf{X})^{-1}\|_2 = \tfrac{\lambda_{\max}}{\lambda_{\min}}, \\ &\kappa_F(\mathbf{X}^\intercal\mathbf{X}) = \|(\mathbf{X}^\intercal\mathbf{X})\|_F \cdot \|(\mathbf{X}^\intercal\mathbf{X})^{-1}\|_F = \left(\left(\textstyle\sum_{i=1}^{p}\lambda_i^2\right)\left(\textstyle\sum_{i=1}^{p}\lambda_i^{-2}\right)\right). \end{aligned} \tag{11}$$

3.2 Polynomial Regression Models

Consider the Polynomial Linear Regression (PLR) model:

$$y_t = \beta_0 + \textstyle\sum_{k=1}^{p}\beta_k x_t^k + u_t, \;\; t \in \mathbb{N}. \tag{12}$$

Using the following matrix notation (Seber and Lee, 2003, p. 165):

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ 1 & x_3 & x_3^2 & \cdots & x_3^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{pmatrix}}_{\mathbf{u}}, \tag{13}$$

as p increases, the $(\mathbf{X}^\intercal\mathbf{X})$ becomes ill-conditioned and the problem becomes worse as the sample size (n) increases; see Montgomery et al. (2012, p. 226). The instability of $(\mathbf{X}^\intercal\mathbf{X})$ is likely to render volatile the estimators:

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}, \quad s^2=\tfrac{1}{n-p}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^\intercal(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}), \tag{14}$$

of $(\boldsymbol{\beta},\sigma^2)$, and give rise to misleading *t-statistics* and *p-values* since $\widehat{Cov(\widehat{\boldsymbol{\beta}})}=s^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$.

Example. To illustrate how the ill-conditioning of $(\mathbf{X}^\intercal\mathbf{X})$ will induced instability in the estimated coefficients and their standard errors, consider the following simplistic but suggestive example with artificial numbers:

$$\mathbf{X}:=\begin{pmatrix} 1.00 & 4.00 \\ 2.01 & 8.00 \\ 4.02 & 16.00 \end{pmatrix}, \ \mathbf{y}:=\begin{pmatrix} 4.0 \\ 2.2 \\ 4.4 \end{pmatrix}, \quad (\mathbf{X}^\intercal\mathbf{X})=\begin{pmatrix} 21.2005 & 84.4 \\ 84.4 & 336.0 \end{pmatrix}, \tag{15}$$

$\kappa_F(\mathbf{X}^\intercal\mathbf{X})=15949024.65$ and $\kappa_2(\mathbf{X}^\intercal\mathbf{X})=15949022.65$, indicating that the two condition numbers are very large and almost identical. In light of the latter, the Frobenius condition number $\kappa_F(\mathbf{X}^\intercal\mathbf{X})$ will be used in what follows since it involves all the eigenvalues. The OLS estimators $\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ and $Var(\widehat{\boldsymbol{\beta}})=\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$ yield:

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}=\begin{pmatrix} 42000.0 & -10550.0 \\ -10550.0 & 2650.0625 \end{pmatrix}\begin{pmatrix} 26.11 \\ 104.0 \end{pmatrix}=\begin{pmatrix} -580.0 \\ 146.0 \end{pmatrix},$$

$$\sqrt{Var(\widehat{\beta}_1)}=204.9\sigma, \ \sqrt{Var(\widehat{\beta}_2)}=51.5\sigma.$$

What are the potential consequences of $(\mathbf{X}^\intercal\mathbf{X})$ being ill-conditioned?

Change 1: Changing the first element of $\mathbf{X}$ from 1.00 to 1.01. The impact on the OLS estimates is that they change signs:

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}=\begin{pmatrix} 42000.0 & -10555.0 \\ -10555.0 & 2652.575 \end{pmatrix}\begin{pmatrix} 26.15 \\ 104.0 \end{pmatrix}=\begin{pmatrix} 580.0 \\ -145.45 \end{pmatrix},$$
$$\sqrt{Var(\widehat{\beta}_1)}=204.9\sigma, \ \sqrt{Var(\widehat{\beta}_2)}=51.5\sigma.$$

Change 2: Changing the first element of $\mathbf{X}$ from 1.00 to 1.02. The result is that the OLS estimates change magnitudes dramatically, and so does the $(\mathbf{X}^\intercal\mathbf{X})^{-1}$, which also affects the significance of (β_1, β_2):

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}=\begin{pmatrix} 4666.667 & -1173.333 \\ -1173.333 & 295.0125 \end{pmatrix}\begin{pmatrix} 26.19 \\ 104.0 \end{pmatrix}=\begin{pmatrix} 193.3 \\ -48.3 \end{pmatrix},$$

$$\sqrt{Var(\widehat{\beta}_1)}=68.3\sigma, \ \sqrt{Var(\widehat{\beta}_2)}=17.2\sigma.$$

Change 3: Changing the first element of $\mathbf{y}$ from 4.0 to 4.5. The impact on the OLS estimates is that their magnitudes change significantly:

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 42000.0 & -10550.0 \\ -10550.0 & 2650.0625 \end{pmatrix} \begin{pmatrix} 26.61 \\ 106.0 \end{pmatrix} = \begin{pmatrix} -680.0 \\ 171.13 \end{pmatrix},$$
$$\sqrt{Var(\widehat{\beta}_1)}=204.9\sigma, \ \sqrt{Var(\widehat{\beta}_2)}=51.5\sigma.$$

3.3 Numerical vs. statistical measures

The traditional econometric account of near-collinearity (Greene, 2011, pp. 129-130) has ignored the important distinction between the numerical and the statistical perspectives. The key difference stems from the fact that numerical measures of ill-conditioning, such as $\kappa(\mathbf{X}^\intercal\mathbf{X})$, pertain to the particular numbers in $(\mathbf{X}^\intercal\mathbf{X})$, irrespective of whether they denote statistical data or arbitrary values, and its effects are evaluated in terms of upper bounds on the volatility of the numerical values of $\widehat{\boldsymbol{\beta}}$ to small changes in the data $(\mathbf{y},\mathbf{X})$; see Spanos and McGuirk (2002). This means that $\kappa(\mathbf{X}^\intercal\mathbf{X})$ has nothing to do with the validity of the model assumptions ([1]-[5]), in contrast to the statistical measures of near-collinearity that are often framed in terms of the sample correlation coefficients among the regressors. In general, moving from numerical measures of ill-conditioning to statistical measures based on sample correlations is not straightforward because any form of statistical misspecification is likely to undermine the reliability of all statistical measures.

For instance, $(\mathbf{y},\mathbf{X})$ are often transformed into sample correlations by taking mean deviations from the sample averages $(\overline{y},\ \overline{x}_1,...,\overline{x}_p)$, which is inappropriate when the data exhibit t-heterogeneity. A case in point is (15), where it is clear that the values of the regressors are trending and thus $\overline{x}_i$ is an inconsistent estimator of $E(X_i)$. This implies that the estimation of correlation $Corr(X_{1t},X_{2t})$ using:

$$\widehat{Corr(X_{1t},X_{2t})}=\frac{\frac{1}{n}\sum_{t=1}^{n}(X_{1t}-\overline{X}_1)(X_{2t}-\overline{X}_2)}{\sqrt{\left[\frac{1}{n}\sum_{t=1}^{n}(X_{1t}-\overline{X}_1)^2\right]\left[\frac{1}{n}\sum_{t=1}^{n}(X_{2t}-\overline{X}_2)^2\right]}}=\frac{84.4}{\sqrt{(21.2005)(336)}}=.9999,$$

will be statistically untrustworthy (spurious). Trending means will also undermine other measures of near-collinearity, such as Variance Inflation Factors (VIFs), $VIF_i=[1/(1-R^2_{[i]})],\ i=2,3,...,k,$ where $R^2_{[i]}$ is the estimated squared multiple correlation coefficient of the auxiliary regression of x_{it} on all the other regressors $\mathbf{x}_{(i)t}$:

$$x_{it}=\alpha_0+\boldsymbol{\alpha}^\intercal_{(i)}\mathbf{x}_{(i)t}+v_{it},\ v_{it}\backsim\mathsf{NIID}(0,\sigma_i^2),\ i=2,3,...,p. \tag{16}$$

Indeed, in the case where the regressors include trend polynomials terms $(t,t^2,...,t^p)$ a constant mean assumption is false by definition, and (16) make no statistical sense when $x_{it}=t^i$, $i=1,2,...,p$.

To account for t-heterogeneity one can use deterministic trend polynomials, which should be treated differently from proper regressors by defining a different sub-matrix

($\mathbf{D}$) separate from the proper regressors, say $\mathbf{X}_1$:

$$\mathbf{y}=\mathbf{D}\boldsymbol{\delta}+\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{u}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}, \tag{17}$$

in an obvious notation. One can estimate $\boldsymbol{\beta}$ using OLS, but including $\mathbf{D}$ in the $\mathbf{X}$ matrix can give rise to highly misleading inference results, especially goodness of fit measures such as the R^2. To get reliable inferential results from computer packages, a practitioner should estimate the formulation in (17) using a two stage procedure.

Stage 1: eliminate $\mathbf{D}$ using the Frisch and Waugh (1933) result based on the projection matrix $\mathbf{M_D}=\mathbf{I}-\mathbf{D}\left(\mathbf{D}^\top\mathbf{D}\right)^{-1}\mathbf{D}^\top=\mathbf{I}-\mathbf{P_D}$, $\mathbf{M_D}\mathbf{M_D}=\mathbf{M_D}$, $\mathbf{M}_\mathbf{D}^\top=\mathbf{M_D}$ and $\mathbf{M_D}\mathbf{D}=\mathbf{0}$. Premultiplying (17) by $\mathbf{M_D}$ yields the transformed specification:

$$\mathbf{M_D}\mathbf{y}=\mathbf{M_D}\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{M_D}\mathbf{u}. \tag{18}$$

Stage 2: estimate the modified form (18) that gives rise to:

$$\widehat{\boldsymbol{\beta}}_1=\left(\mathbf{X}_1^\top\mathbf{M_D}\mathbf{X}_1\right)^{-1}\mathbf{M_D}\mathbf{y},\ s^2=\tfrac{\widehat{\mathbf{u}}^\top\mathbf{M_D}\widehat{\mathbf{u}}}{n-p-k},\ \widehat{\mathbf{u}}=(\mathbf{M_D}\mathbf{y}-\mathbf{M_D}\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1). \tag{19}$$

Note that $\mathbf{M_D}\mathbf{y}=\mathbf{y}^*$ and $\mathbf{M_D}\mathbf{X}_1=\mathbf{X}^*$ yield the mean deviations of $\mathbf{y}$ and $\mathbf{X}_1$ from the trend polynomial ($\mathbf{D}$), giving rise to the correct goodness-of-fit formula:

$$R^2_\mathbf{D}=1-\tfrac{\widehat{\mathbf{u}}^\top\mathbf{M_D}\widehat{\mathbf{u}}}{\mathbf{y}^\top\mathbf{M_D}\mathbf{y}}, \tag{20}$$

and $\left(\mathbf{X}_1^\top\mathbf{M_D}\mathbf{X}_1\right)$ provides the basis for the sample correlation for $\mathbf{X}_{1t}$:

$$\mathbf{R}^*_X:=\mathbf{Q}^{-1}\left(\mathbf{X}_1^\top\mathbf{M_D}\mathbf{X}_1\right)\mathbf{Q}^{-1},\ \mathbf{Q}:=\text{diag}(\sqrt{\textstyle\sum_{t=1}^n(x^*_{2t})^2},...,\sqrt{\textstyle\sum_{t=1}^n(x^*_{pt})^2}). \tag{21}$$

This amounts to 'detrending' the data before the correlations among the regressors $\mathbf{X}_1$ are evaluated. Note that $\mathbf{D}$ could also include other deterministic terms, such as seasonal and 0:1 dummy variables.

4 Trend polynomials and ill-conditioning

4.1 Ordinary trend polynomials

A common case where the near-collinearity problem arises is when fitting ordinary trend polynomials using least-squares. The use of these polynomials offers a generic way to 'capture' a variety of different forms of t-heterogeneity. In such a context, the practitioner can construct a data matrix $\mathbf{T}$ whose column space consists of $(t^0, t^1, t^2, t^3, ..., t^p)$, for $t=1, 2, ..., n$. What is important to note is that from a purely numerical analysis perspective, $\mathbf{T}$ is a special case of a Vandermonde matrix $\mathbf{V}=[\nu_{i,j}]$, where $\nu_{i,j}=\nu_i^{j-1}$, for $i=1, 2, ..., n,\ j=1, 2, ..., p$, and $(\mathbf{T}^\intercal\mathbf{T})$ is a Hankel matrix $\mathbf{H}=[h_{i,j}]_{i,j=1}^p$, where $h_{i,j}=h_{i+1,j-1}=h_{i+j-2}$. It is well known that both Vandermonde and Hankel matrices tend to be ill-conditioned, even for low degree polynomials p; see Tyrtyshinikov (1994) and Pan (2016).

Table 3 reports the Frobenius norm condition numbers computed using (11), for different sample sizes (n) and degree of the polynomials (p). The results show that the Frobenius norm condition numbers grow exponentially with both the n and p. In point of fact, the ill-conditioning appears to be particularly severe even for n=100 and p=3 since $\kappa_F\left(\mathbf{X}^\intercal\mathbf{X}\right)=2.57\times10^{12}$. A closer look at the diagonal values of the $(\mathbf{X}^\intercal\mathbf{X})$ matrix indicates that there are large differences among those values.

Table 3 - Ordinary Trend Polynomials: $\kappa_F\left(\mathbf{X}^\intercal\mathbf{X}\right)$			
n	p=3	p=5	p=7
n=100	2.57×10^{12}	4.27×10^{20}	7.05×10^{28}
n=250	5.84×10^{14}	3.44×10^{24}	1.89×10^{34}

In light of that, one might conjecture that a possible way to reduce the near-collinearity problem will be to change $(t^0, t, t^2, t^3, ..., t^p)$ using *monotonic transformations*:

(i) scale their range by the sample size n,

(ii) use logarithmic transformations,

(iii) scale their range to lie within the interval $[-1, 1]$ (Seber and Lee, 2003, p. 166):

$$x_i^*=\frac{(2x_i-x_{[n]}-x_{[1]})}{(x_{[n]}-x_{[1]})},\ \ i=1, 2, ..., n, \tag{22}$$

where $x_{[n]}=\max(x_1, x_2, ..., x_n)$, $x_{[1]}=\min(x_1, x_2, ..., x_n)$.

The results in table 4 suggest that the transformed data matrix $\mathbf{X}$ has a substantially reduced norm condition number. For example, the norm condition number for n=250 and p=7 is reduced by $\times10^{30}$ when the range of t:=$(1, 2, ..., n)$ is scaled to lie within $[-1, 1]$. In spite of that, the transformed data matrices remain Vandermonde in form, showing signs of $(\mathbf{X}^\intercal\mathbf{X})$ being ill-conditioned. This indicates that the scaling has a significant effect on tempering the increase in n, since the norm condition numbers remain relatively stable as the sample size increases, but does little to curtail the effect of increasing p.

Table 4 - Scaled Ordinary Trend Polynomials: $\kappa_F\left(\mathbf{X}^\intercal\mathbf{X}\right)$					
Scaling	Range	n	p=3	p=5	p=7
x_i^*	$[-1, 1]$	n=100	72.07	1.99×10^{3}	5.86×10^{4}
		n=250	73.71	2.06×10^{3}	6.10×10^{4}
$\frac{x_i}{n}$	$(0, 1]$	n=100	1.61×10^{4}	1.59×10^{7}	1.67×10^{10}
		n=250	1.58×10^{4}	1.54×10^{7}	1.59×10^{10}
$\ln(x_i)$	$[0, \infty)$	n=100	1.01×10^{6}	7.74×10^{9}	1.83×10^{14}
		n=250	4.88×10^{6}	4.60×10^{10}	1.18×10^{15}
$\ln(\frac{x_i}{n})$	$(-\infty, 0]$	n=100	1.16×10^{4}	1.02×10^{8}	1.50×10^{12}
		n=250	1.31×10^{4}	1.50×10^{8}	2.93×10^{12}

In addition, it is interesting to examine the magnitude of the determinants of the $(\mathbf{X}^\intercal\mathbf{X})$ matrix presented in table 5. The determinants of the (unscaled) ordinary trend polynomials suggest that $(\mathbf{X}^\intercal\mathbf{X})$ is not close to being singular, but the determinants of the polynomials scaled by n approximate zero for higher degree of polynomials.

Table 5 - Determinants: $\det(\mathbf{X}^\intercal\mathbf{X})$ when scaled by n				
Scaling	n	p=3	p=5	p=7
x_i	n=100	1.65×10^{25}	5.31×10^{54}	2.65×10^{95}
	n=250	3.85×10^{31}	1.13×10^{69}	8.00×10^{120}
$\frac{x_i}{n}$	n=100	16.50	5.31×10^{-6}	2.65×10^{-17}
	n=250	645.67	1.31×10^{-3}	4.15×10^{-14}

These results suggest that $\det(\mathbf{X}^\intercal\mathbf{X})\simeq0$ is not a good measure of the near singularity of $(\mathbf{X}^\intercal\mathbf{X})$. For instance, for $\mathbf{A}_n=\text{diag}(10^{-1},...,10^{-1})$, $\det(\mathbf{A}_n)=10^{-n}$, but $\kappa_2(\mathbf{A}_n)=1$; see Golub and Van Loan (2013, p. 89).

4.2 Orthogonal polynomials

A common way to reduce near-collinearity is to replace the ordinary with orthogonal polynomials, $\{\phi_k(.),\ k=0,1,2,...\}$; see Seber and Lee (2003, p. 166). It is important, however, to distinguish between continuous and discrete polynomials.

Definition 1 *A sequence of continuous polynomials* $\{\varphi_k(x), k=0,1,2,...\}$ *is orthogonal with respect to a continuous weight function* $w(x)\geq0$ *on the interval* (a,b) *if:*

$$\int_a^b w(x)\varphi_i(x)\varphi_j(x)dx=c_i\delta_{ij},\ \ i,j=0,1,2,..., \tag{23}$$

where $c_i=\int_a^b w(x)\left[\varphi_i(x)\right]^2dx$, *and* $\delta_{ij}=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$.

Definition 2 *A sequence of discrete polynomials* $\{\wp_k(t), k=0,1,2,...,N\}$ *is called orthogonal with respect to a positive weight* w_t *on the countable set* $\mathbb{T}$*:*

$$\sum\nolimits_{t\in\mathbb{T}} w_t\wp_i(t)\wp_j(t)=c_i\delta_{ij},\ \ i,j=0,1,2,...,N, \tag{24}$$

where $c_i=\sum\nolimits_{t\in\mathbb{T}} w_t\left[\wp_i(t)\right]^2$, *and* $\delta_{ij}=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$.

For these polynomials the $\mathbf{X}$ matrix in (13) takes the form of a Vandermonde-like matrix whose column space consists of the p-th degree of orthogonal polynomials:

$$\mathbf{X}=([\phi_i(x_t)]_{i,t}),\ \ x\in(a,b),\ \ i=0,1,...,p,\ \ t=1,2,...,n, \tag{25}$$

where x_t is replaced with t, for discrete polynomials.

The choice between continuous and discrete polynomials depends on both whether t is assigned discrete or continuous values and the range of such values. In the context of modeling t-heterogeneity using trend polynomials with $t{=}1, 2, ..., n$, one can combine a discrete t with a continuous or a discrete polynomial. As shown below, however, the continuous/discrete choice does not address the ill-conditioning problem; hence the need for rescaling.

To evaluate the ill-conditioning of $(\mathbf{X}^\intercal\mathbf{X})$ using various continuous and discrete orthogonal polynomials (see Appendix A), we report their norm condition numbers for different sample sizes (n) and degree of polynomials (p). Table 6 reports the lowest condition numbers of the polynomials among the monotonic transformations (i)-(ii) for $t{:=}(1, 2, ..., n)$ - scaled by the sample size and logarithmic transformations. For comparison purposes, table 7 reports the condition numbers for the same polynomials using the scaling in (iii) - equally spaced ordering on the interval $[-1, 1]$. The scaling for continuous polynomials ensures that the interval of orthogonality is pertinent. In addition, lower bounds for the condition numbers are calculated for polynomials whose parameters can take different values (in square brackets). Note that the Chebyshev, Legendre and Gegenbauer are special cases of the Jacobi polynomials.

Surprisingly, the norm condition numbers in table 6 are similar to the ones of the scaled ordinary polynomials in table 4. Specifically, the norm condition numbers grow exponentially with the degree of the polynomial, but remain stable with respect to increases to the sample size. In a nutshell, all the orthogonal polynomials show signs of $(\mathbf{X}^\intercal\mathbf{X})$ being ill-conditioned, whereas the only polynomials that indicate marginal improvements are the Jacobi. These results indicate that orthogonality itself is not sufficient enough to overcome the near-collinearity problem.

By contrast, the results in table 7 point out that the norm condition numbers of some orthogonal polynomials are significantly reduced when the scaling in (iii) is applied. Of particular interest is the fact that the continuous polynomials whose interval of orthogonality is $[-1, 1]$, including the Jacobi, Gegenbauer, Chebyshev, and Legendre, appear to be well-conditioned. On the other hand, all the discrete polynomials, as well as continuous polynomials whose orthogonality is outside the interval $[-1, 1]$, like the Hermite, show signs of ill-conditioning. Given these results, the main conclusion seems to be that one way to reduce the near-collinearity problem when fitting trend polynomials using least-squares is to combine orthogonality and scaling of the time ordering. That is, the practitioner should use an equally spaced transformation of the original time ordering $t{\in}\mathbb{N}{:=}(1, 2, ..., n)$, on the interval $[-1, 1]$, in conjuction with continuous orthogonal polynomials whose interval of orthogonality is $[-1, 1]$.

4.3 From orthogonal to orthonormal polynomials

Of course one may argue that a more direct way to avoid near-collinearity when using trend polynomials is to use the Gram-Schmidt (G-S) orthonormal polynomials resulting from transforming the sequence $\left\{t^k, k{=}0,1,2,...\right\}$ to recursively generate:

$$v_0{=}1, \quad v_k{=}t_k{-}\sum_{l=0}^{k-1}\eta_l v_l, \;\; \eta_l{=}\tfrac{(t_k,v_l)}{(v_l,v_l)}, \;\; \widetilde{v}_k{=}\tfrac{v_k}{\|v_k\|}, \;\; k{=}0,1,2,..., \tag{26}$$

the polynomials $\{\widetilde{v}_k, k{=}0,1,2,...\}$, rendered orthonormal with $\kappa_F\left(\mathbf{X}^\intercal\mathbf{X}\right){=}1$; see Golub and Van Loan (2013, pp. 254-255). The question that naturally arises is whether the G-S orthonormal polynomials render the previous discussion pertaining to other polynomials redundant? The answer is clearly no because in addition to dealing with near-collinearity issues, the different polynomials might be more appropriate for different types of inference.

For instance, for estimation (point and interval) and testing purposes the G-S polynomials have a clear advantage, but for forecasting and simulation purposes other orthogonal polynomials might have a clear advantage because they have explicit functional forms that do not need to be re-evaluated as the G-S polynomials when the sample size changes. Hence, in cases where the inference of interest calls for explicit functional forms for the polynomials, one is better off by choosing among the continuous orthogonal polynomials defined over the interval $[-1,1]$, such as the Jacobi; see table 7. Yet, if one is explicitly interested in long-term forecasting, it might be preferable to use less rigid polynomials such as splines; see Ruppert et al. (2003).

Table 6 - Scaled Orthogonal polynomials: $\kappa_F\left(\mathbf{X}^\intercal\mathbf{X}\right)$

Panel A: Continuous Orthogonal Polynomials

Polynomial	Sample size	p=3	p=5	p=7
Jacobi $[\alpha;\beta]$	n=100	10.16 $[.95; 7.63]$	90.7 $[1.49; 11.65]$	9.00×10^2 $[2.02; 16.16]$
	n=250	10.09 $[.97; 7.62]$	89.7 $[1.52; 11.63]$	8.86×10^2 $[2.07; 16.13]$
Gegenbauer $[\lambda]$	n=100	7.33×10^3 $[.82]$	6.09×10^6 $[.79]$	5.74×10^9 $[.78]$
	n=250	7.12×10^3 $[.82]$	5.77×10^6 $[.80]$	5.29×10^9 $[.79]$
Chebyshev (1st kind)	n=100	1.61×10^4	1.76×10^7	2.01×10^{10}
	n=250	1.59×10^4	1.71×10^7	1.93×10^{10}
Chebyshev (2nd kind)	n=100	8.04×10^3	7.23×10^6	7.39×10^9
	n=250	7.76×10^3	6.75×10^6	6.60×10^9
Legendre	n=100	1.15×10^4	1.04×10^7	1.04×10^{10}
	n=250	1.13×10^4	1.01×10^7	9.94×10^9
Hermite	n=100	1.17×10^5	1.47×10^{10}	1.15×10^{16}
	n=250	1.16×10^5	1.44×10^{10}	1.11×10^{16}
Laguerre	n=100	3.32×10^5	3.08×10^{10}	1.35×10^{16}
	n=250	2.31×10^5	1.29×10^{10}	2.89×10^{15}

Panel B: Discrete Orthogonal Polynomials

Polynomial	Sample size	p=3	p=5	p=7
Charlier $[\alpha]$	n=100	3.19×10^2 $[52.92]$	3.82×10^3 $[49.25]$	4.60×10^4 $[46.94]$
	n=250	3.16×10^2 $[132.72]$	4.21×10^3 $[123.83]$	7.03×10^4 $[119.61]$
Chebyshev $[N]$	n=100	2.15×10^4 $[6]$	2.47×10^7 $[6]$	7.58×10^{16} $[8]$
	n=250	9.36×10^4 $[7]$	1.69×10^8 $[7]$	6.16×10^{14} $[8]$
Krawtchouk $[\theta; N]$	n=100	25.86 $[.87; 4]$	6.00×10^3 $[.68; 5]$	3.11×10^7 $[.44; 7]$
	n=250	15.62 $[.85; 5]$	1.44×10^3 $[.87; 5]$	1.62×10^6 $[.58; 7]$
Meixner $[\beta; c]$	n=100	35.32$[2.69; .91]$	$7.47\times10^2[4.34; .85]$	$2.17\times10^4[5.08; .81]$
	n=250	29.76$[2.97; .96]$	$4.76\times10^2[5.10; .93]$	$8.97\times10^3[6.06; .91]$

The table presents the Frobenius norm condition numbers of various scaled continuous and discrete orthogonal trend polynomials, for different sample sizes (n) and degree of polynomials (p). The polynomials are scaled by re-scaling their original time ordering, $t = 1, 2, ..., n$, either (i) by the sample size, n, or (ii) by a logarithmic transformation. The reported norm condition numbers are the lowest among the two scalings. For polynomials whose parameters can take different values, the lower bounds of the norm condition numbers are calculated. The parameter values corresponding to the lower bounds are given in square brackets.

Table 7 - Orthogonal Polynomials over $[-1,1]$: $\kappa_F(\mathbf{X}^{\intercal}\mathbf{X})$

Panel A: Continuous Orthogonal Polynomials

Polynomial	Sample size	p=3	p=5	p=7
Jacobi $[\alpha;\beta]$	n=100	4.59 [.31; .31]	7.85 [.29; .29]	11.75 [.27; .27]
	n=250	4.54 [.32; .32]	7.59 [.31; .31]	11.05 [.29; .29]
Gegenbauer $[\lambda]$	n=100	4.59 [.81]	7.85 [.79]	11.75 [.77]
	n=250	4.54 [.82]	7.59 [.81]	11.05 [.79]
Chebyshev (1st kind)	n=100	8.06	13.01	17.63
	n=250	8.41	14.21	20.36
Chebyshev (2nd kind)	n=100	5.49	10.39	17.24
	n=250	5.33	9.65	15.11
Legendre	n=100	9.51	17.32	26.43
	n=250	9.75	17.96	27.56
Hermite	n=100	1.93×10^2	8.26×10^5	2.26×10^{10}
	n=250	2.02×10^2	8.87×10^5	2.44×10^{10}
Laguerre	Range of values of t lie outside the interval $[-1,1]$			

Panel B: Discrete Orthogonal Polynomials

Polynomial	Sample size	p=3	p=5	p=7
Charlier $[\alpha]$	n=100	3.02×10^4 [.86]	1.30×10^{10} [1.54]	2.35×10^{16} [4.34]
	n=250	3.10×10^4 [.86]	1.36×10^{10} [1.53]	1.74×10^{16} [4.47]
Chebyshev $[N]$	n=100	2.17×10^7 [4]	1.14×10^{17} [6]	8.23×10^{28} [8]
	n=250	2.21×10^7 [4]	1.18×10^{17} [6]	6.02×10^{27} [8]
Krawtchouk $[\theta;N]$	n=100	3.61×10^3 [.01; 3]	7.95×10^7 [.01; 5]	8.81×10^{12} [.01; 7]
	n=250	3.73×10^3 [.01; 3]	8.35×10^7 [.01; 5]	9.38×10^{12} [.01; 7]
Meixner $[\beta;c]$	n=100	1.98×10^3 [.58; .06]	2.17×10^7 [.09; .01]	2.57×10^{12} [.01; .01]
	n=250	2.02×10^3 [.57; .06]	2.25×10^7 [.09; .01]	2.71×10^{12} [.01; .01]

The table presents the Frobenius norm condition numbers of various continuous and discrete orthogonal trend polynomials whose original time ordering, $t=1,2,...,n$, is scaled to lie within the interval $[-1,1]$, for different sample sizes (n) and degree of polynomials (p). For polynomials whose parameters can take different values, the lower bounds of the norm condition numbers are calculated. The parameter values corresponding to the lower bounds are given in square brackets.

5 An empirical example: Aggregate Portfolios

The grouping of assets into portfolios in asset pricing modeling is viewed as a way to reduce the errors-in-variables problem in such data, and thus improve the precision of the estimated betas; see Blume (1970), Friend and Blume (1970), Black et al. (1972). There is, however, a potential side-effect when the original data exhibit different forms of t-heterogeneity. This means that averaging individual such asset returns could generate complicated forms of t-heterogeneity for the portfolio. This is especially problematic for portfolios composed of a very large number of assets. Let us illustrate how one can capture the t-heterogeneity in such data using continuous orthogonal polynomials defined over the interval $[-1, 1]$, or G-S orthonormal polynomials.

This empirical example uses the Fama and French (2015) data from July 1963 to December 2015. The portfolio employed is the (smallest size/lowest profitability) portfolio which includes all NYSE, AMEX, and NASDAQ stocks whose market capitalization and operating profitability are respectively within the bottom quintile of NYSE stocks. Since most AMEX and NASDAQ stocks are smaller in market capitalization than the NYSE stocks, this portfolio (on average), contains approximately 1058 stocks, a lot more than portfolios constructed with similar characteristics.

After thorough misspecification testing, the following heterogeneous Student's t/ heteroskedastic dynamic LR model of the Capital Asset Pricing Model (CAPM) is estimated in order to account for the chance regularity patterns exhibited by the observed data in hand:

$$\begin{gathered} y_t = \alpha + \sum_{i=1}^{12} \delta_{1i} t^i + \sum_{j=2}^{12} \delta_{2j} d_{jt} + \beta_1 x_{1t} + \beta_2 x_{2t} + \boldsymbol{\gamma}_1^\top \mathbf{Z}_{t-1} + \varepsilon_t, \\ (\varepsilon_t | \mathbf{Z}_{t-1}) \backsim \mathsf{St}(0, \sigma^2(t)), \;\; \sigma^2(t) = \gamma_0 + [\mathbf{W}_t - \boldsymbol{\mu}(t)]^\top \mathbf{Q}^{-1} [\mathbf{W}_t - \boldsymbol{\mu}(t)], \end{gathered} \tag{27}$$

where y_t is the return of the (smallest size/lowest profitability) portfolio for period t (figure 1); $x_{1t}=R_{mt}$ is the return on the value-weighted market portfolio (figure 2); $x_{2t}=R_{ft}$ is the return on the risk-free asset (figure 3); $t^i:=(t, t^2, ..., t^{12})$, denotes the polynomials of {Ordinary over $[-1, 1]$, Meixner by n, Hermite over $[-1, 1]$, Jacobi over $[-1, 1]$, G-S orthonormal}; $d_{jt}:=(d_{2t}, d_{3t}, ..., d_{12t})$ are the monthly dummy variables for the months of February through December; $\mathbf{Z}_t:=(y_t, \mathbf{X}_t)$, $\mathbf{X}_t:=(x_{1t}, x_{2t})$, $\mathbf{W}_t:=(\mathbf{X}_t, \mathbf{Z}_{t-1})$, $\boldsymbol{\mu}(t)=E(\mathbf{W}_t)$, $\mathbf{Q}=Cov(\mathbf{W}_t)$. Note that $\sigma^2(t)$ is both heteroskedastic and heterogeneous, stemming from the Multivariate Student's t distribution.

Table 8 presents the results from estimating the same overall model in (27) by including the different types of trend polynomials. The inflated estimated coefficients and standard errors (in brackets) for the cases of ordinary over $[-1, 1]$, Meixner scaled by the sample size, and Hermite over $[-1, 1]$ polynomials indicate the presence of serious near-collinearity, which is also confirmed by the inflated norm condition numbers. It is increasingly difficult for the practitioner to assess the relative contribution of each polynomial in terms of statistical significance since their p-values (in square brackets) are highly misleading. The picture becomes much clearer when the G-S orthonormal polynomials or the Jacobi over $[-1, 1]$ polynomials are used to capture the t-heterogeneity in the data. As can be seen from table 8, the absence of near-collinearity for these polynomials provides trustworthy evidence for the significance of the higher degree trend polynomials up to degree 12. It is important to emphasize that the choice of the highest degree p of the trend polynomial estimated is based exclusively on statistical adequacy grounds, and not on goodness-of-fit as in Akaike type model selection procedures; see Spanos (2010) for a discussion on this. That is, one would continue to increase p until the residuals do not have lingering t-heterogeneity and the coefficient of the term with the largest p is statistically significant. In this case, p=12 is not an unreasonably large value because the data are monthly, but it could have been smaller on adequacy grounds.

The results in table 8 confirm that neither the scaling of the time ordering nor the orthogonality property alone are sufficient enough to overcome the near-collinearity problem when fitting trend polynomials using least-squares. Yet, the most efficient way to adequately address the near-collinearity problem is to use the G-S orthonormal polynomials, or to combine continuous orthogonal polynomials whose interval of orthogonality is $[-1, 1]$, with scaling the time ordering to lie within the same interval.

The importance of this result stems from the fact that establishing the statistical adequacy of the estimated model, secures the reliability of inferences based on it. In the case of the empirical multi-factor models, like the Fama-French models (see Fama and French, 1993, 2015), these polynomials secure the statistical adequacy of the estimated models which can provide a sound basis for evaluating the significance of the additional factors. Capturing the t-heterogeneity in this data, gives rise to trustworthy results when posing substantive questions of interest to the data, including which additional factors are significant or not. In contrast, when the t-heterogeneity in the data is not captured, the statistical significance of the added factors is likely to be untrustworthy.

Table 8 - Modeling heterogeneity is asset returns

y_t	Ordinary	Meixner [β=9.02; c=.91]	Hermite	Jacobi [α=.29; β=.29]	G-S orthonormal
c	2.108[.001] (0.636)	3.756[.000] (0.637)	-13.620[.844] (69.155)	3.098[.000] (0.538)	3.101[.000] (0.522)
t	-2.804[.157] (1.980)	-0.967[.047] (0.485)	57.022[.929] (639.717)	-0.577[.001] (0.166)	-12.696[.000] (3.555)
t^2	32.352[.008] (12.233)	-0.878[.205] (0.691)	- 91.199 [.929] (1017.265)	0.231[.199] (0.180)	3.134[.392] (3.660)
t^3	-14.478[.627] (29.815)	5.474[.000] (1.515)	42.768 [.987] (2645.153)	0.049[.714] (0.134)	1.058[.657] (2.379)
t^4	-187.952[.125] (122.399)	8.345[.000] (1.958)	- 41.418 [.990] (3155.453)	-0.393[.001] (0.122)	-7.050[.001] (2.142)
t^5	157.738[.330] (161.963)	0.567[.835] (2.724)	-115.527[.978] (4173.752)	-0.432[.004] (0.151)	-4.093[.098] (2.473)
t^6	467.668[.338] (488.179)	-10.053[.005] (3.571)	106.080[.978] (3854.339)	0.424[.003] (0.141)	4.671[.039] (2.264)
t^7	-363.972[.331] (374.098)	-11.637[.006] (4.191)	- 38.292 [.979] (1450.451)	0.588[.000] (0.153)	11.727[.000] (2.458)
t^8	-606.016[.510] (918.897)	-6.738[.098] (4.061)	13.097[**NaN**] (**NaN**)	-0.705[.000] (0.154)	-11.304[.000] (2.389)
t^9	296.251[.440] (383.424)	4.189[.539] (6.820)	202.963[**NaN**] (**NaN**)	0.589[.000] (0.142)	7.569[.001] (2.207)
t^{10}	423.236[.604] (815.902)	1.128[.837] (5.471)	-160.951[**NaN**] (**NaN**)	-0.281[.095] (0.168)	-1.454[.550] (2.429)
t^{11}	-71.542[.619] (143.884)	6.819[.118] (4.352)	138.265[**NaN**] (**NaN**)	-0.445[.001] (0.137)	-7.134[.001] (2.217)
t^{12}	-130.003[.637] (275.466)	-2.961[.730] (8.569)	-94.774[**NaN**] (**NaN**)	0.626[.000] (0.145)	9.956[.000] (2.301)
d_2	-2.457[.000] (0.467)	-2.512[.000] (0.464)	-2.405[.000] (0.459)	-2.462[.000] (0.471)	-2.458[.000] (0.471)
d_3	-2.469[.000] (0.446)	-2.328[.000] (0.441)	-2.394[.000] (0.440)	-2.478[.000] (0.445)	-2.473[.000] (0.445)
d_4	-3.815[.000] (0.483)	-3.767[.000] (0.477)	-3.868[.000] (0.478)	-3.932[.000] (0.482)	-3.930[.000] (0.481)
d_5	-2.533[.000] (0.477)	-2.596[.000] (0.472)	-2.528[.000] (0.468)	-2.607[.000] (0.473)	-2.603[.000] (0.473)
d_6	-2.635[.000] (0.489)	-2.715[.000] (0.484)	-2.484[.000] (0.482)	-2.757[.000] (0.485)	-2.753[.000] (0.485)
d_7	-3.334[.000] (0.501)	-3.273[.000] (0.493)	-3.259[.000] (0.491)	-3.535[.000] (0.500)	-3.533[.000] (0.500)
d_8	-3.234[.000] (0.483)	-3.302[.000] (0.483)	-3.106[.000] (0.467)	-3.569[.000] (0.485)	-3.559[.000] (0.485)
d_9	-2.296[.000] (0.493)	-2.249[.000] (0.484)	-2.254[.000] (0.477)	-2.522[.000] (0.480)	-2.522[.000] (0.479)
d_{10}	-4.150[.000] (0.512)	-4.278[.000] (0.509)	-4.191[.000] (0.512)	-4.281[.000] (0.505)	-4.277[.000] (0.505)
d_{11}	-3.586[.000] (0.493)	-3.654[.000] (0.488)	-3.605[.000] (0.485)	-3.814[.000] (0.485)	-3.811[.000] (0.485)
d_{12}	-3.174[.000] (0.473)	-3.319[.000] (0.470)	-3.143[.000] (0.462)	-3.344[.000] (0.474)	-3.337[.000] (0.474)
x_{1t}	1.254[.000] (0.029)	1.240[.000] (0.029)	1.259[.000] (0.029)	1.249[.000] (0.028)	1.249[.000] (0.028)
x_{2t}	-5.587[.091] (3.305)	-5.657[.087] (3.298)	-4.733[.154] (3.317)	-4.967[.124] (3.225)	-4.957[.125] (3.229)
y_{t-1}	0.103[.019] (0.044)	0.096[.029] (0.044)	0.128[.002] (0.042)	0.078[.070] (0.043)	0.079[.067] (0.043)
x_{1t-1}	0.174[.010] (0.067)	0.187[.005] (0.067)	0.154[.021] (0.066)	0.203[.002] (0.065)	0.203[.002] (0.065)
x_{2t-1}	2.590[.448] (3.414)	2.856[.403] (3.412)	1.887[.581] (3.421)	2.686[.421] (3.333)	2.683[.422] (3.336)
$\kappa\left(\mathbf{X}^{\intercal}\mathbf{X}\right)_F$	3.37×10^{8}	1.16×10^{7}	1.12×10^{15}	20.07	13

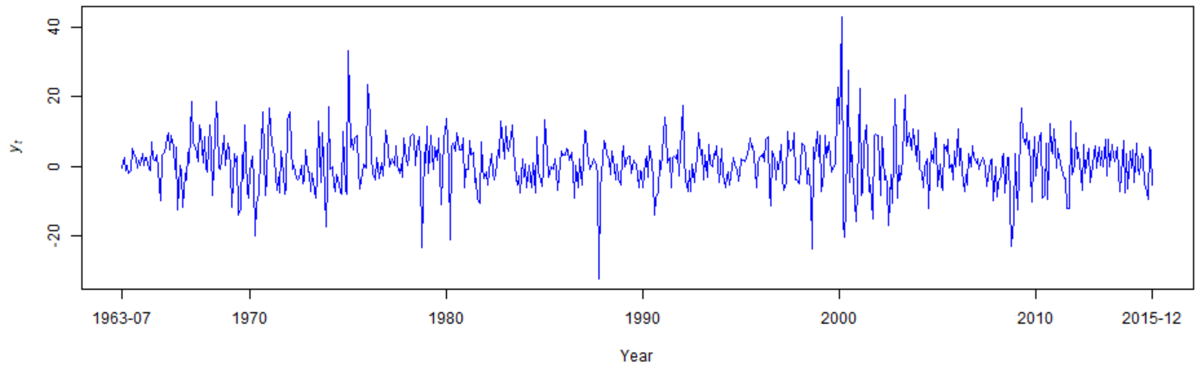

Figure 1: t-plot of portfolio return y_t; July 1963 to December 2015

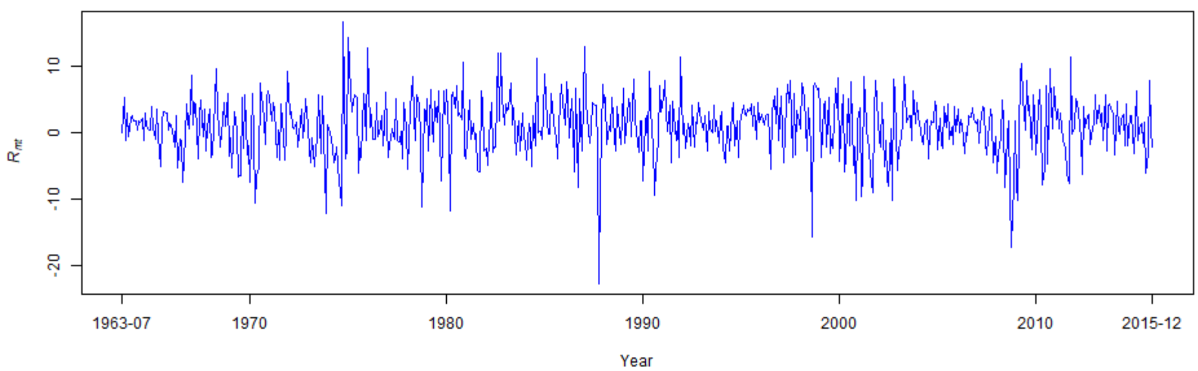

Figure 2: t-plot of market return R_{mt}; July 1963 to December 2015

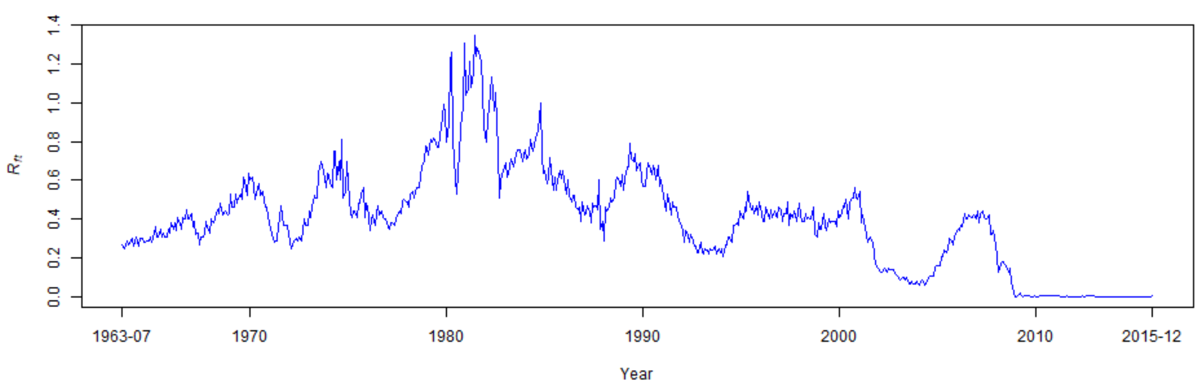

Figure 3: t-plot of risk-free return R_{ft}; July 1963 to December 2015

6 Summary and conclusions

The paper has focused on evaluating a number of different ways one can address the problem of modeling t-heterogeneity in data as well as the ill-conditioning problem that arises when using trend polynomials to account for such heterogeneity in the context of $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u}$ when $\mathbf{X}$ includes trend polynomials based on the ordering $t=1,2,...,n$. In such cases, the problem of ill-conditioning can be mitigated using a twofold strategy. The first is to orthogonalize the polynomial terms $t, t^2, ..., t^p$. This led us to the evaluation of several continuous and discrete orthogonal polynomials, such as the Chebyshev, Hermite, Jacobi, Laguerre, the Krawtchouk, Charlier, and Meixner. The second is a scaling of the trend ordering ($t=1,2,...,n$). This led us to evaluate different forms of scaling and concluded that the best is the one which yields approximately equally spaced discrete values over the interval $[-1,1]$. The ill-conditioning can be effectively circumvented by combining the orthogonality with this particular scaling.

The discussion in the paper can be summarized by a few recommendations to the practitioner. First, any form of t-heterogeneity exhibited by the data must be accounted for using generic methods, such as trend polynomials, because ignoring such systematic information in the data will give rise to statistically spurious results. Second, the simplest way to avoid serious near-collinearity problems when fitting trend polynomials of degree lower than 7 or so, is simply to re-scale the original time ordering using (22) to confine its range of values within $[-1,1]$. Third, the recommendation for trend polynomials of degree greater than 7 or so is the G-S orthonormal polynomials because they are particularly useful for estimation (point and interval) and testing purposes. Fourth, for forecasting and simulation purposes the practitioner might prefer an explicit functional form for the orthogonal polynomials, and not a sequence of values for a fixed sample size, as given by the G-S polynomials. For such purposes the recommendation is to choose among certain continuous orthogonal polynomials, such as the Jacobi, specified over the interval $[-1,1]$. Fifth, if the practitioner is only interested in long-term forecasting, it might be preferable to use less rigid polynomials such as splines; see Ruppert et al. (2003). Sixth, to avoid misleading statistical measures of near-collinearity, it is important to treat the deterministic polynomials $\mathbf{D}:=(t, t^2, ..., t^p)$ differently from the other regressors $\mathbf{X}_t$ and apply the Frisch-Waugh result before certain statistical measure. This amounts to 'detrending' the data using the relevant mean before the second moments of $(y_t, \mathbf{X}_t)$ and the R^2 are evaluated.

References

[1] Black, F., Jensen, M. C., and Scholes, M. S. (1972), "The Capital Asset Pricing Model: Some Empirical Tests," pp. 79-121 in *Studies in the Theory of Capital Markets*, edited by Jensen, M.C., Praeger, NY.

[2] Blume, M.E. (1970), "Portfolio Theory: A Step Toward its Practical Application," *Journal of Business*, 43(2), 152-173.

[3] Choi, I. (2015), *Almost All about Unit Roots: Foundations, Developments, and Applications*, Cambridge University Press, Cambridge.

[4] Dunkl, C.F., and Xu, Y. (2014), *Orthogonal Polynomials of Several Variables*, 2nd ed., Encyclopedia of Mathematics and its Applications, Vol. 155, Cambridge University Press, Cambridge.

[5] Fama, E.F., and French, K.R. (1993), "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33(1), 3-56.

[6] Fama, E.F., and French, K.R. (2015), "A Five-Factor Asset Pricing Model," *Journal of Financial Economics*, 116(1), 1-22.

[7] Friend, I., and M. Blume, M. (1970), "Measurement of Portfolio Performance Under Uncertainty," *American Economic Review*, 60(4), 561-575.

[8] Frisch, R., and Waugh, F.V. (1933), "Partial Time Regressions as Compared with Individual Trends," *Econometrica*, 1(4), 387-401.

[9] Gautschi, W. (1990), "How (Un)stable Are Vandermonde Systems," *Asymptotic and Computational Analysis,* 124, 193-210.

[10] Gautschi, W. (2004), *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press, NY.

[11] Gautschi, W., and Inglese, G. (1988), "Lower Bounds for the Condition Number of Vandermonde Matrices," *Numerische Mathematik*, 52(3), 241-250.

[12] Golub, G.H., and Van Loan, C.F. (2013), *Matrix Computations*, 4th ed., Johns Hopkins, Baltimore.

[13] Granger, C.W.J., and Newbold, P. (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics*, 2(2), 111-120.

[14] Greene, W.H. (2011), *Econometric Analysis*, 7th ed., Prentice Hall, New Jersey.

[15] Hooker, R.H. (1901), "Correlation of the Marriage-Rate with Trade," *Journal of the Royal Statistical Society*, 64(3), 485-492.

[16] Hooker, R.H. (1905), "On the Correlation of Successive Observations," *Journal of the Royal Statistical Society*, 68(4), 696-703.

[17] Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012), *Introduction to Linear Regression Analysis*, 5th ed., Wiley, NJ.

[18] Moore, H.L. (1914), *Economic Cycles: Their Law and Cause*, Macmillan, London.

[19] Morgan, M.S. (1990), *The History of Econometric Ideas*, Cambridge University Press, Cambridge.

[20] Pan, V.Y. (2016), "How Bad are Vandermonde Matrices," *SIAM Journal on Matrix Analysis and Applications*, 37(2), 676-694.

[21] Phillips, P.C.B. (1986), "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33(3), 311-340.

[22] Phillips, P.C.B. (2005), "Challenges of Trending Time Series Econometrics," *Mathematics and Computers in Simulation*, 68, 401-416.

[23] Ruppert, D., Wand, M.P., and Carroll R.J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge.

[24] Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis*, 2nd ed., Wiley, NY.

[25] Spanos, A. (2019), *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*, 2nd ed., Cambridge University Press, Cambridge.

[26] Spanos, A. (2010), "Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification," *Journal of Econometrics*, 158(2), 204-220.

[27] Spanos, A., and McGuirk, A. (2002), "The Problem of Near-multicollinearity Revisited: Erratic vs. Systematic Volatility," *Journal of Econometrics*, 108(2), 365-393.

[28] Stigler, S.M. (1986), *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard University Press.

[29] Tyrtyshnikov, E.E. (1994), "How bad are Hankel matrices?," *Numerische Mathematik*, 67(2), 261-269.

[30] Yule, G.U. (1903), "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, 2, 121-134.

[31] Yule, G.U. (1926) "Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series," *Journal of the Royal Statistical Society*, 89(1), 1-63.

7 Appendix A: Orthogonal Polynomials

For additional information, see Gautschi (2004), Dunkl and Xu (2014).

Definition 3 *The Jacobi polynomials* $P_k^{(\alpha,\beta)}(x)$*, for parameters* $\alpha, \beta{>}-1$*, are a set of orthogonal polynomials with weight function* $w(x;\alpha,\beta){=}(1{-}x)^{\alpha}(1{+}x)^{\beta}$ *on* $x{\in}[-1,1]$*.*

$$P_k^{(\alpha,\beta)}(x){=}\tfrac{(-1)^k}{2^k k!}(1{-}x)^{-\alpha}(1{+}x)^{-\beta}\left(\tfrac{d}{dx}\right)^k\left[(1{-}x)^{\alpha+k}(1{+}x)^{\beta+k}\right],\ \ k{\geq}0.$$

The recurrence relation of the Jacobi polynomials is:

$$\begin{aligned}
&2(k+1)(k+\alpha+\beta+1)(2k+\alpha+\beta)P_{k+1}^{(\alpha,\beta)}(x){=}\\
&{=}[(2k{+}\alpha{+}\beta{+}1)(2k{+}\alpha{+}\beta{+}2)(2k{+}\alpha{+}\beta)x{+}\alpha^2{-}\beta^2]P_k^{(\alpha,\beta)}(x){-}\\
&-2(k+\alpha)(k+\beta)(2k+\alpha+\beta+2)P_{k-1}^{(\alpha,\beta)}(x),\quad k{\geq}1,\\
&P_0^{(\alpha,\beta)}(x){=}1,\quad P_1^{(\alpha,\beta)}(x){=}\tfrac{1}{2}(\alpha+\beta+2)x+\tfrac{1}{2}(\alpha-\beta).
\end{aligned}$$

Definition 4 *The Gegenbauer polynomials* $C_k^{(\lambda)}(x)$*, for parameter* $\lambda{>}-\frac{1}{2}$*, are a set of orthogonal polynomials with weight function* $w(x;\lambda){=}(1-x^2)^{\lambda-\frac{1}{2}}$ *on* $x{\in}[-1,1]$*.*

$$C_k^{(\lambda)}(x){=}\tfrac{(-1)^k}{2^k(\lambda+\frac{1}{2})_k}(1-x^2)^{\frac{1}{2}-\lambda}\left(\tfrac{d}{dx}\right)^k(1-x^2)^{\lambda+k-\frac{1}{2}},\ \ k{\geq}0.$$

The recurrence relation of the Gegenbauer polynomials is:

$$\begin{gathered}
(k+1)C_{k+1}^{(\lambda)}(x){=}2(k{+}\lambda)xC_k^{(\lambda)}(x){-}(k{+}2\lambda{-}1)C_{k-1}^{(\lambda)}(x),\quad k{\geq}1,\\
C_0^{(\lambda)}(x){=}1,\quad C_1^{(\lambda)}(x){=}2\lambda x.
\end{gathered}$$

The Gegenbauer polynomials can be viewed as a special case of the Jacobi polynomials with $\alpha{=}\beta{=}\lambda-\frac{1}{2}$.

Definition 5 *The Chebyshev polynomials of the first kind* $T_k(x)$ *are a set of orthogonal polynomials with weight function* $w(x){=}(1-x^2)^{-\frac{1}{2}}$ *on* $x{\in}[-1,1]$*.*

$$T_k(x){=}\tfrac{(-1)^k 2^k k!}{(2k)!}(1-x^2)^{\frac{1}{2}}\left(\tfrac{d}{dx}\right)^k(1-x^2)^{k-\frac{1}{2}},\ \ k{\geq}0.$$

Definition 6 *The Chebyshev polynomials of the second kind* $U_k(x)$ *are a set of orthogonal polynomials with weight function* $w(x){=}(1-x^2)^{\frac{1}{2}}$ *on* $x{\in}[-1,1]$*.*

$$U_k(x){=}\tfrac{(-1)^k 2^k (k+1)!}{(2k+1)!}(1-x^2)^{-\frac{1}{2}}\left(\tfrac{d}{dx}\right)^k(1-x^2)^{k+\frac{1}{2}},\ \ k{\geq}0.$$

The recurrence relation of the Chebyshev polynomials is:

$$y_{k+1}=2xy_k-y_{k-1}, \quad k\geq 1,$$
$$y_0=1, \quad y_1(T_k(x))=x, \quad y_1(U_k(x))=2x.$$

The Chebyshev polynomials of the first and second kind can be viewed as special cases of the Jacobi polynomials with $\alpha=\beta=-\frac{1}{2}$ and $\alpha=\beta=\frac{1}{2}$, respectively. Further, the Chebyshev polynomials of the first and second kind can also be viewed as special cases of the Gegenbauer polynomials with $\lambda=0$ and $\lambda=1$, respectively.

Definition 7 *The Legendre polynomials $P_k(x)$ are a set of orthogonal polynomials with weight function $w(x)=1$ on $x\in[-1,1]$.*

$$P_k(x)=\frac{(-1)^k}{2^k k!}\left(\frac{d}{dx}\right)^k (1-x^2)^k, \ k\geq 0.$$

The recurrence relation of the Legendre polynomials is:

$$(k+1)P_{k+1}(x)=(2k+1)xP_k(x)-kP_{k-1}(x), \quad k\geq 1,$$
$$P_0(x)=1, \quad P_1(x)=x.$$

The Legendre polynomials can be viewed as a special case of the Jacobi polynomials with $\alpha=\beta=0$. Further, the Legendre polynomials can also be viewed as a special case of the Gegenbauer polynomials with $\lambda=\frac{1}{2}$.

Definition 8 *The Hermite polynomials $H_k(x)$ are a set of orthogonal polynomials with weight function $w(x)=e^{-x^2}$ on $x\in\mathbb{R}$.*

$$H_k(x)=(-1)^k e^{x^2}\left(\frac{d}{dx}\right)^k e^{-x^2}, \ k\geq 0.$$

The recurrence relation of the Hermite polynomials is:

$$H_{k+1}(x)=2xH_k(x)-2kH_{k-1}(x), \quad k\geq 1,$$
$$H_0(x)=1, \quad H_1(x)=2x.$$

Definition 9 *The Laguerre polynomials $L_k^{(\alpha)}(x)$, for parameter $a>-1$, are a set of orthogonal polynomials with weight function $w(x;a)=e^{-x}x^{\alpha}$ on $x\in\mathbb{R}_+$.*

$$L_k^{(\alpha)}(x)=\frac{e^x x^{-\alpha}}{k!}\left(\frac{d}{dx}\right)^k \left(e^{-x}x^{k+\alpha}\right), \ k\geq 0.$$

The recurrence relation of the Laguerre polynomials is:

$$(k+1)L_{k+1}^{(\alpha)}(x)=(2k+\alpha+1-x)L_k^{(\alpha)}(x)-(k+\alpha)L_{k-1}^{(\alpha)}(x), \quad k\geq 1,$$
$$L_0^{(\alpha)}(x)=1, \quad L_1^{(\alpha)}(x)=\alpha+1-x.$$

The classical Laguerre polynomials corresponds to $\alpha=0$.

Definition 10 *The Discrete Chebyshev polynomials (also known as Gram polynomials) $D_k^{(N)}(t)$, for parameter $N \in \mathbb{N}$, are a set of orthogonal polynomials with weight function $w(t)=1$ on $t=0,1,...,N-1$.*

$$D_k^{(N)}(t)=\Delta^k \frac{(t-k+1)_k (t-N-k+1)_k}{k!}, \;\; k=0,1,2,...,N-1.$$

The recurrence relation of the Discrete Chebyshev polynomials is:

$$(k+1)D_{k+1}^{(N)}(t)=2(2k+1)(t-\tfrac{1}{2}(N-1))D_k^{(N)}(t)-t(N^2-t^2)D_{k-1}^{(N)}(t), \;\; k=0,1,2,...,N-1,$$
$$D_0^{(N)}(t)=1, \quad D_1^{(N)}(t)=2t-N+1.$$

Definition 11 *The Charlier polynomials $C_k^{(\alpha)}(t)$, for parameter $\alpha>0$, are a set of orthogonal polynomials with weight function $w(t;\alpha)=\frac{e^{-\alpha}\alpha^t}{t!}$ on $t \in \mathbb{N}_0$.*

$$C_k^{(\alpha)}(t)=\frac{t!}{\alpha^t} \triangle^k \left[\frac{\alpha^{t-k}}{(t-k)!}\right], \;\; k\geq 0.$$

The recurrence relation of the Charlier polynomials is:

$$\alpha C_{k+1}^{(\alpha)}(t)=(k+\alpha-t)C_k^{(\alpha)}(t)-kC_{k-1}^{(\alpha)}(t), \;\; k\geq 0,$$
$$C_0^{(\alpha)}(t)=1, \quad C_1^{(\alpha)}(t)=-\frac{1}{\alpha}t.$$

Definition 12 *The Krawtchouk polynomials $K_k^{(\theta,N)}(t)$, for parameters $0<\theta<1$ and $N \in \mathbb{N}$, are a set of orthogonal polynomials with weight function $w(t;\theta,N)=\begin{pmatrix} N \\ t \end{pmatrix} \theta^t(1-\theta)^{N-t}$ on $t=0,1,...,N$.*

$$K_k^{(\theta,N)}(t)=\frac{(-1)^k t!(N-t)!}{k!\theta^t(1-\theta)^{N-t}} \triangle^k \left[\frac{\theta^t(1-\theta)^{N-t+k}}{(t-k)!(N-t)!}\right], \;\; k=0,1,2,...,N.$$

The recurrence relation of the Krawtchouk polynomials is:

$$(k+1)K_{k+1}^{(\theta,N)}(t)=[t-(k+\theta(N-2k))]K_k^{(\theta,N)}(t)-(N-k+1)\theta(1-\theta)K_{k-1}^{(\theta,N)}(t), \;\; k=0,1,2,...,N,$$
$$K_0^{(\theta,N)}(t)=1, \quad K_1^{(\theta,N)}(t)=t-\theta N.$$

Definition 13 *The Meixner polynomials (also known as discrete Laguerre polynomials) $M_k^{(\beta,c)}(t)$, for parameters $\beta>0$ and $0<c<1$, are a set of orthogonal polynomials with weight function $w(t;\beta,c)=\frac{c^t(\beta)_t}{t!}$ on $t \in \mathbb{N}_0$.*

$$M_k^{(\beta,c)}(t)=\frac{t!}{(\beta)_t}c^{-t-k} \triangle^k \left[\frac{c^t(\beta)_t}{(t-k)!}\right], \;\; k\geq 0.$$

The recurrence relation of the Meixner polynomials is:

$$(k+\beta)cM_{k+1}^{(\beta,c)}(t)=[(c-1)t+(1+c)k+\beta c]M_k^{(\beta,c)}(t)-kM_{k-1}^{(\beta,c)}(t), \;\; k\geq 0,$$
$$M_0^{(\beta,c)}(t)=1, \quad M_1^{(\beta,c)}(t)=\frac{c-1}{\beta c}t+1.$$