The Stata Journal (2018)
18, Number 3, pp. 585–617

Estimating dynamic common-correlated effects in Stata

Jan Ditzen
Centre for Energy Economics Research and Policy
and
Spatial Economics and Econometrics Centre
Heriot-Watt University
Edinburgh, UK
j.ditzen@hw.ac.uk

Abstract. In this article, I introduce a new command, `xtdcce2`, that fits a dynamic common-correlated effects model with heterogeneous coefficients in a panel with a large number of observations over cross-sectional units and time periods. The estimation procedure mainly follows Chudik and Pesaran (2015b, *Journal of Econometrics* 188: 393–420) but additionally supports the common correlated effects estimator (Pesaran, 2006, *Econometrica* 74: 967–1012), the mean group estimator (Pesaran and Smith, 1995, *Journal of Econometrics* 68: 79–113), and the pooled mean group estimator (Pesaran, Shin, and Smith, 1999, *Journal of the American Statistical Association*, 94: 621–634). `xtdcce2` allows heterogeneous or homogeneous coefficients and supports instrumental-variable regressions and unbalanced panels. The cross-sectional dependence test is automatically calculated and presented in the estimation output. Small-sample time-series bias can be corrected by "half-panel" jackknife correction or recursive mean adjustment. I carry out a simulation to prove the estimator's consistency.

Keywords: st0536, xtdcce2, xtcd2, parameter heterogeneity, dynamic panels, cross-section dependence, common correlated effects, pooled mean group estimator, mean group estimator, instrumental variables, ivreg2

1 Introduction

Estimating panels with heterogeneous coefficients in a panel with a large dimension of observations over cross-sectional units (N) and time periods (T) became standard in the last years because of seminal work in theoretical econometrics (Pesaran and Smith 1995; Pesaran, Shin, and Smith 1999). Heterogeneous slopes allow the researcher to identify effects for each cross-section separately. Concurrently, the theoretical literature on how to account for unobserved dependence between cross-sectional units evolved (Pesaran 2006; Chudik and Pesaran 2015b). Not accounting for unobserved dependence between cross-sectional units causes the error term to be autocorrelated and leads to biased ordinary least-squares (OLS) regression results.

In this article, I introduce a new command, `xtdcce2`, that combines these two strands of the literature and allows for mean group (MG) estimations in a dynamic panel with

dependence between cross-sectional units.[1] `xtdcce2` obtains MG estimates in two steps: First, the coefficients of interest of each cross-sectional unit are estimated and therefore allow for heterogeneous slopes. Second, the unit-specific estimates are averaged across all groups. `xtdcce2` controls for cross-sectional dependence by adding cross-sectional averages and lags as proposed by Pesaran (2006) and Chudik and Pesaran (2015b).[2] Furthermore, it tests for weak cross-sectional dependence in the error terms and allows for instrumental-variable (IV) estimation. Additionally, `xtdcce2` allows correction for small-sample time-series bias by using the "half-panel" jackknife correction method or the recursive mean adjustment as proposed by Chudik and Pesaran (2015b).

`xtdcce2` differs in several ways from the existing estimation procedures for common correlated effects (CCE) in a heterogeneous panel. Compared with `xtmg` (Eberhardt 2012), it allows the consistent estimation of a dynamic panel by adding lags of the cross-sectional means. Moreover, it can constrain coefficients to be homogeneous across all units. Additionally, it supports unbalanced panels. Compared with the `xtpmg` command (Blackburne and Frank 2007), `xtdcce2` avoids maximum likelihood estimations, offering the possibility to fit models including endogenous independent variables. Hence, the main novelties within the setting of `xtpmg` and `xtmg` are the inclusion of a test for cross-sectional dependence, small T bias-correction methods, and the support for IV regressions. IV regressions benefit from the `ivreg2` package. Possible applications for an IV estimation are endogenous spatial lags, which are instrumented by exogenous measures such as distance, other variables, or higher-order spatial lags. Furthermore, adding cross-sectional means implies accounting for unobserved heterogeneity across units.

The `xtdcce2` package includes `xtcd2`, which tests for weak cross-sectional dependence (henceforth the CD test) as proposed by Pesaran (2015) and Chudik and Pesaran (2015a). Two other commands, `xtcd` by Eberhardt (2011) and `xtcsd` by De Hoyos and Sarafidis (2006), made the CD test already available in Stata. The novelty of `xtcd2` is the support of unbalanced panels, the possibility to test any variable for cross-sectional dependence, and the option to plot the cross correlations as a kernel density plot.

The remainder of this article is structured as follows: The next two sections briefly introduce dynamic common-correlated effects (DCCE) and testing for cross-sectional dependence. I then explain the syntax, options, and stored results of `xtdcce2` and `xtcd2`. I close with examples for an empirical application and a comparison of regression results obtained by `xtdcce2` with results from estimation procedures already available in Stata, followed by a simulation to prove the estimator's consistency.

1. I describe `xtdcce2`, version 1.31. For updates, use `search xtdcce2` in Stata or see the author's webpage.
2. Chudik and Pesaran (2015a) comprehensively overview the literature on (dynamic) common correlated effects (CCE), while Chudik and Pesaran (2015b) focus on dynamic common-correlated effects (DCCE). In the following, I cite Pesaran (2006) for CCE, while I cite Chudik and Pesaran (2015b) for DCCE, even though both are found in Chudik and Pesaran (2015a).

2 CCE estimators

Assume the following equation with heterogeneous coefficients (Pesaran 2006),

$$\begin{aligned} y_{it} &= \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it} \\ u_{it} &= \boldsymbol{\gamma}_i' \mathbf{f}_t + e_{it} \end{aligned} \tag{1}$$

where $\mathbf{f}_t$ is an unobserved common factor, $\boldsymbol{\gamma}_i$ a heterogeneous factor loading, and α_i a unit-specific fixed effect.[3] e_{it} is a cross-section unit-specific independent and identically distributed (IID) error term. The heterogeneous coefficients are randomly distributed around a common mean such that $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{v}_i$, $\mathbf{v}_i \sim \text{IID}(\mathbf{0}, \boldsymbol{\Omega}_v)$, where $\boldsymbol{\Omega}_v$ is the variance–covariance matrix. Pesaran (2006) shows that (1) can be consistently estimated by approximating the unobserved common factors with cross-sectional averages $\overline{\mathbf{x}}_t$ under strict exogeneity of $\mathbf{x}_{it}$. This estimator is commonly known as the CCE estimator. The underlying idea of the CCE estimator is to eliminate asymptotically the differential effects of unobserved common factors by cross-sectional averages as the cross-sectional dimension approaches infinity (Pesaran 2006, 969).

The estimator was proved consistent under a variety of further assumptions on the error term (see Chudik, Pesaran, and Tosetti [2011] and Kapetanios, Pesaran, and Yamagata [2011]). In empirical applications, it was used, for example, in Eberhardt, Helmers, and Strauss (2013), Bond and Eberhardt (2013), McNabb and LeMay-Boucher (2014), and Gundlach and Paldam (2016). The CCE estimator was made available in Stata by Eberhardt's (2012) `xtmg` command.

However, the CCE estimator is consistent only in nondynamic panels (Chudik and Pesaran 2015b; Everaert and Groote 2016). In a dynamic panel such as

$$y_{it} = \alpha_i + \lambda_i y_{i,t-1} + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it} \tag{2}$$

where the idiosyncratic errors u_{it} are cross-sectionally weakly dependent and $E(\lambda_i) = \lambda$, the lagged dependent variable is no longer strictly exogenous. Therefore, the estimator becomes inconsistent. Chudik and Pesaran (2015b) show that the estimator gains consistency if the floor of $\sqrt[3]{T}$ lags of the cross-section averages is added for both the dependent variables and the strictly exogenous variables. Let's denote the number of lags by $p_T = \lfloor \sqrt[3]{T} \rfloor$. The equation to be estimated is

$$y_{it} = \alpha_i + \lambda_i y_{i,t-1} + \boldsymbol{\beta}_i' \mathbf{x}_{it} + \sum_{l=0}^{p_T} \boldsymbol{\delta}_{il}' \overline{\mathbf{z}}_{t-l} + e_{it}$$

where $\overline{\mathbf{z}_t} = (\overline{y}_{t-1}, \overline{\mathbf{x}}_t)$. Consider λ_i and $\boldsymbol{\beta}_i$ as stacked into $\boldsymbol{\pi}_i = (\lambda_i, \boldsymbol{\beta}_i)$; then the MG estimates are

$$\widehat{\boldsymbol{\pi}}_{\text{MG}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{\pi}}_i$$

3. Unlike in Pesaran (2006), the unit-specific fixed effect is kept and not partialed out. See discussion in section 4.6.

$\widehat{\boldsymbol{\pi}}_i$ and $\widehat{\boldsymbol{\pi}}_{\text{MG}}$ are consistently estimated if $(N, T, p_T) \xrightarrow{j} \infty$ such that $p_T^3/T \rightarrow \varrho_1, 0 < \varrho < \infty$ and $N/T \rightarrow \varrho_2, \varrho_2 > 0$ and under full rank of the factor loadings (Chudik and Pesaran 2015b). The requirements for consistency on the two estimators can be interpreted for each separately. The unit-specific estimates can be obtained from a simple regression on a single cross-section unit. Therefore, the requirement for consistency is $T \rightarrow \infty$. A relative expansion rate for N and T is not required. The number of cross-sectional lags is restricted to maintain a sufficient number of degrees of freedom. Therefore, the requirement on the number of lags is necessary. For consistency of the MG estimates, N and T grow jointly to infinity ($[N, T] \xrightarrow{j} \infty$). The cross-sectional dimension approaches infinity because of the heterogeneous coefficients. The time dimension grows to reduce the time series because of the lagged dependent variable. For a more in-depth discussion, see section 3 and 3.1 in Chudik and Pesaran (2015a).

Under these assumptions, the asymptotic variance for the MG estimates is consistently estimated by

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\pi}}_{\text{MG}}) = N^{-1}\widehat{\boldsymbol{\Sigma}}_{\pi} = \frac{1}{N(N-1)}\sum_{i=1}^{N}(\widehat{\boldsymbol{\pi}}_i - \widehat{\boldsymbol{\pi}}_{\text{MG}})(\widehat{\boldsymbol{\pi}}_i - \widehat{\boldsymbol{\pi}}_{\text{MG}})'$$

The MG estimates have the following asymptotic distribution (Chudik and Pesaran 2015b):

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_{\text{MG}} - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{MG}})$$

Pesaran (2006) considers a pooled version of the CCE estimator, with the constraint $\boldsymbol{\pi}_i = \boldsymbol{\pi}\ \forall i$. In the case of equal weights to all observations, $w_i = 1/N\ \forall i$, the CCE pooled estimator for $\boldsymbol{\pi}$, denoted as $\widehat{\boldsymbol{\pi}}_P$, collapses to a simple OLS estimator. Everaert and Groote (2016) show that a CCE pooled estimator even in a dynamic panel is consistent as long as $(N, T) \Rightarrow \infty$. Following Pesaran (2006), a nonparametric variance estimator for $\widehat{\boldsymbol{\pi}}_P$ is given by

$$\widehat{\text{AVar}}(\widehat{\boldsymbol{\pi}}_P) = \frac{1}{N}\widehat{\boldsymbol{\Psi}}^{*-1}\widehat{\mathbf{R}}^{*}\widehat{\boldsymbol{\Psi}}^{*-1} \tag{3}$$

with

$$\widehat{\mathbf{R}}^{*} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i}{T}\right)(\widehat{\boldsymbol{\pi}}_i - \widehat{\boldsymbol{\pi}}_{mg})(\widehat{\boldsymbol{\pi}}_i - \widehat{\boldsymbol{\pi}}_{mg})'\left(\frac{\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i}{T}\right)$$

where $\widetilde{\mathbf{X}}_i$ are the explanatory variables with the cross-sectional averages partialed out and

$$\widehat{\boldsymbol{\Psi}}^{*} = \sum_{i=1}^{N}\frac{1}{N}\left(\frac{\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i}{T}\right)$$

The asymptotic distribution for the pooled estimator is

$$\sqrt{N}(\widehat{\boldsymbol{\pi}}_P - \boldsymbol{\pi}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_P)$$

The pooled MG estimator (Pesaran, Shin, and Smith 1999) can be seen as an intermediate between a pure pooled estimation (homogeneous coefficients) and an MG estimation (heterogeneous coefficients). The assumptions of the pooled MG estimator are that regressors have a homogeneous long-run effect and a heterogeneous short-run effect on the dependent variable. Equation (2) is transformed into an error-correction model such that

$$\Delta y_{it} = \phi_i(y_{it-1} - \boldsymbol{\theta}_i'\mathbf{x}_{it}) + \alpha_i + \boldsymbol{\beta}_i'\Delta\mathbf{x}_{it} + u_{it}$$

$\phi_i = (1-\alpha_i)$ is the error-correction speed of the adjustment parameter and is expected to be negative; $(y_{i,t-1} - \boldsymbol{\theta}_i'\mathbf{x}_{it})$ is the error-correction term. $\boldsymbol{\theta}_i = \boldsymbol{\beta}_i/\phi_i$ is the long-run coefficient and assumed to be homogeneous, while $\boldsymbol{\beta}_i$ captures short-term dynamics and is heterogeneous across units.[4] Pesaran, Shin, and Smith (1999) propose to estimate the long-run coefficients by maximum likelihood and the short-run coefficients by OLS. The estimator is consistent as long as the disturbances are independently distributed across all individuals and time periods with a zero mean and a variance strictly larger than zero.

The MG estimate and the variance of the short-run coefficients are

$$\widehat{\boldsymbol{\delta}}_{\text{MG}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\delta}}_i, \qquad \widehat{\text{Var}}(\widehat{\boldsymbol{\delta}}_{\text{MG}}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\delta}}_i - \widehat{\boldsymbol{\delta}}_{\text{MG}}\right)^2$$

where $\boldsymbol{\delta}_i = (\alpha_i, \boldsymbol{\beta}_i)$.

The MG and the pooled MG estimator in the static and dynamic versions rely on large N and T. The literature on small-sample time-series bias corrections in dynamic heterogeneous panels is somewhat scarce, so Chudik and Pesaran (2015b) focus on "half-panel" jackknife and recursive mean-adjustment bias-correction methods. Neither requires knowledge of the error-factor structure, and they can be applied to the MG estimates.[5] The MG estimate of the "half-panel" jackknife bias-corrected CCE estimator is

$$\widetilde{\boldsymbol{\pi}}_{\text{MG}} = 2\widehat{\boldsymbol{\pi}}_{\text{MG}} - \frac{1}{2}\left(\widehat{\boldsymbol{\pi}}^a_{\text{MG}} + \widehat{\boldsymbol{\pi}}^b_{\text{MG}}\right)$$

where $\widehat{\boldsymbol{\pi}}^a_{\text{MG}}$ is the MG estimate of the first half $(t = 1, \dots, T_i/2)$ of the panel and $\widehat{\boldsymbol{\pi}}^b_{\text{MG}}$ of the second half $(t = T_i/2+1, \dots, T_i)$ of the panel.

The recursive mean adjustment removes the partial mean from all the variables, meaning

$$\widetilde{\boldsymbol{\omega}}_{it} = \boldsymbol{\omega}_{it} - \frac{1}{t-1}\sum_{s=1}^{t-1}\boldsymbol{\omega}_{is}$$

where $\boldsymbol{\omega}_{it} = (y_{it}, \boldsymbol{x}_{it})$ or any other variable except the constant. In line with Chudik and Pesaran (2015b), the partial mean is lagged by one period to prevent it from being influenced by endogenous observations.

4. This notation follows (1) of Pesaran, Shin, and Smith (1999) with $p = q = 1$.
5. See Chudik and Pesaran (2015b) or Everaert and De Vos (2016).

From the asymptotics of the unit-specific estimates and the MG estimates, restrictions on the dataset arise. The number of cross-sectional units and time periods is assumed to grow at the same rate. In an empirical setting, this can be interpreted as N/T being constant. A dataset with one dimension being large compared with the other would lead to inconsistent estimates even if both dimensions are large in number. For example, a financial dataset on stock market returns on a monthly basis over 30 years ($T = 360$) of 10,000 firms would not be sufficient. While both dimensions can be interpreted individually as large, they do not grow at the same rate, and the ratio would not be constant. Therefore, an estimator relying on fixed T asymptotics and large N would be appropriate. On the other hand, a dataset with, say, $N = 30$ and $T = 34$ would qualify as appropriate if N and T grow at the same rate.[6]

3 Testing for weak cross-sectional dependence

If (1) is estimated without accounting for the error structure, the unobserved common factor and the heterogeneous factor loading remain a part of the error term u_{it}. In this case, u_{it} will be correlated across units (in other words, cross-sectionally dependent), and the error will not be IID anymore. An omitted-variable bias problem occurs if the observed explanatory variables and the unobserved common factors are correlated. In both cases, OLS becomes inconsistent (Everaert and Groote 2016). Chudik, Pesaran, and Tosetti (2011) describe two types of cross-sectional dependence. Following the notations from (1), the error term is weakly cross-sectionally dependent if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |\boldsymbol{\gamma}_i| = 0$$

and strongly cross-sectionally dependent if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |\boldsymbol{\gamma}_i| \geq K > 0.$$

Cross-sectional independence is defined by $\boldsymbol{\gamma}_i = 0\ \forall i$. However, cross-sectional independence is a restrictive assumption for large panels, and only strong cross-sectional dependence poses a problem (Pesaran 2015).

Pesaran (2015) develops a procedure to test for weak cross-sectional dependence. Under the null hypothesis, the error terms are weakly cross-sectionally dependent.[7] The test statistic is

$$\text{CD} = \sqrt{\frac{2T}{N\,(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \widehat{\rho}_{ij} \right)$$

$$\widehat{\rho}_{ij} = \widehat{\rho}_{ji} = \frac{\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt}}{\left(\sum_{t=1}^{T} \widehat{u}_{it}^2\right)^{1/2} \left(\sum_{t=1}^{T} \widehat{u}_{jt}^2\right)^{1/2}}$$

6. Thanks to an anonymous referee for suggesting the two examples.
7. For a more formal derivation of the null hypothesis, see Pesaran (2015).

where $\widehat{\rho}_{ij}$ is the correlation coefficient.[8] In the case of an unbalanced panel, the correlation coefficient is calculated for the common sample

$$\widehat{\rho}_{ij} = \widehat{\rho}_{ji} = \frac{\sum_{t\in T_i\cap T_j}\left(\widehat{u}_{it}-\overline{\widehat{u}}_i\right)\left(\widehat{u}_{jt}-\overline{\widehat{u}}_j\right)}{\left\{\sum_{t\in T_i\cap T_j}\left(\widehat{u}_{it}-\overline{\widehat{u}}_i\right)^2\right\}^{(1/2)}\left\{\sum_{t\in T_i\cap T_j}\left(\widehat{u}_{jt}-\overline{\widehat{u}}_j\right)^2\right\}^{(1/2)}}$$

where

$$\overline{\widehat{u}}_i = \frac{\sum_{t\in T_i\cap T_j}\widehat{u}_{it}}{T_{ij}}, \quad T_{ij} = \#\left(T_i\cap T_j\right)$$

and where $T_i \cap T_j$ are the common periods of unit i and j and $\#\left(T_i\cap T_j\right)$ is the number of common periods. The CD test statistic then becomes

$$\text{CD} = \sqrt{\frac{2}{N\left(N-1\right)}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sqrt{T_{ij}}\widehat{\rho}_{ij}\right)$$

Under the null, the CD test statistic is asymptotically distributed:

$$\text{CD} \sim N(0,1)$$

For a more in-depth discussion, see Pesaran (2015) and Chudik and Pesaran (2015a).

4 The xtdcce2 command

4.1 Syntax

```
xtdcce2 depvar [indepvars] [(varlist2 = varlist_iv)] [if] [in],
    {crosssectional(varlist_cr)|nocrosssectional} [pooled(varlist_p)
    cr_lags(#) ivreg2options(string) e_ivreg2 ivslow noisily lr(varlist_lr)
    lr_options(string) pooledconstant reportconstant noconstant trend
    pooledtrend jackknife recursive nocd showindividual fullsample]
```

Data must be `xtset` before using `xtdcce2`. *depvar*, *indepvars*, *varlist2*, *varlist_iv*, *varlist_cr*, *varlist_p*, and *varlist_lr* may contain time-series operators (for reference, see [U] **11.4.4 Time-series varlists**) and factor variables (for reference, see [U] **11.4.3 Factor variables**). `xtdcce2` requires the `moremata` package by Jann (2005). *varlist2* are the endogenous variables and *varlist_iv* are the instruments.

8. The index for the time periods is omitted for the balanced panel.

4.2 Options

`crosssectional(`*varlist_cr*`)` defines the variables that are included in $\mathbf{z}_t$ and added as lagged cross-sectional averages ($\overline{\mathbf{z}}_{t-l}$) to the equation. The coefficients of the lagged cross-sectional averages are treated as nuisance parameters that have no interpretation and are therefore partialed out.

`crosssectional(_all)` adds all variables from *depvar*, *indepvars*, *varlist2*, *varlist_iv*, and *varlist_lr* as cross-sectional averages. No cross-sectional averages are added if `crosssectional(_none)` is used, which is equivalent to `nocrosssectional`.

`crosssectional()` is required but can be substituted by `nocrosssectional`. Variables in `crosssectional()` may be included in *varlist_p*, *varlist2*, *varlist_iv*, and *varlist_lr*.

`nocrosssectional` suppresses adding cross-sectional averages. Results will be equivalent to the Pesaran and Smith (1995) MG estimator or, if `lr(`*varlist*`)` is specified, to the Pesaran, Shin, and Smith (1999) pooled MG estimator.

`pooled(`*varlist_p*`)` specifies homogeneous coefficients. For these variables, the estimated coefficients are constrained to be equal across all units ($\boldsymbol{\beta}_i = \boldsymbol{\beta}\ \forall\ i$). Variables may occur in *indepvars*, *varlist2*, *varlist_iv*, *varlist_cr*, and *varlist_lr*.

`cr_lags(#)` specifies the number of lagged cross-sectional averages. For example, `cr_lags(2)` includes the contemporaneous cross-sectional averages and the first and second lag of the cross-sectional averages. If not defined, but `crosssectional()` contains *varlist_cr* or `cr_lags(0)`, then only contemporaneous cross-sectional averages are added but no lags.

`xtdcce2` allows IV regression. *varlist_2* specifies endogenous variables, and *varlist_iv* specifies exogenous variables from IV regression using the `ivreg2` command written by Baum, Schaffer, and Stillman (2003, 2007). The use of *varlist_iv* and *varlist_2* requires the prior installation of `ivreg2`.

`ivreg2options(`*string*`)` passes further options onto `ivreg2`; see `ivreg2` for more information.

`e_ivreg2` posts all available results from `ivreg2` in `e()` with prefix `ivreg2_`.

`ivslow` request using `ivreg2` for the calculation of auxiliary regressions rather than a faster `mata` routine. For the calculation of standard errors for pooled coefficients, an auxiliary regression is performed. In this regression, all coefficients are heterogeneous. If option `ivslow` is used, then `xtdcce2` calls `ivreg2` for the auxiliary regression. This is advisable as soon as `ivreg2`-specific options are used, which influences point estimates.

`noisily` shows the output of the wrapped `ivreg2` regression command.

`xtdcce2` is able to fit pooled mean group models (Pesaran, Shin, and Smith 1999), similarly to `xtpmg`.

`lr(`*varlist_lr*`)` specifies the variables to be included in the long-run cointegration vector in addition to the error-correcting speed of the adjustment term. Using the notation from (2) with the error-correction term as $(y_{i,t-1} - \boldsymbol{\theta}_i' \mathbf{x}_{it})$, the option would read `lr(L.y x)`.

`lr_options(`*string*`)` specifies options for the long-run coefficients. Options may be the following:

`nodivide`, where coefficients are not divided by the error-correction speed of the adjustment vector [that is, estimate (5)]; and

`xtpmgnames`, where coefficient names in `e(b)` (or `e(bi)`) and `e(V)` (or `e(Vi)`) match the name convention from `xtpmg`.

`pooledconstant` restricts the constant to be the same across all groups ($\beta_{0,i} = \beta_0, \ \forall i$).

`reportconstant` reports the constant term. If not specified, the constant is treated as a part of the cross-sectional averages and partialed out.

`noconstant` suppresses the constant term.

`trend` adds a linear unit-specific trend, t_i. It may not be combined with `pooledtrend`.

`pooledtrend` adds a linear common trend. It may not be combined with `trend`.

Two methods for small-sample time-series bias correction are supported:

`jackknife` applies the "half-panel" jackknife bias correction for small-sample time-series bias. It may not be combined with `recursive`.

`recursive` applies the recursive mean-adjustment method to correct for small-sample time-series bias. It may not be combined with `jackknife`.

`nocd` suppresses calculation of the CD test statistic.

`showindividual` reports cross-sectional unit-specific estimates in output.

`fullsample` uses the entire sample available for calculation of cross-sectional averages. Any observations that are lost because of lags will be included in calculating the cross-sectional averages but are not included in the estimation itself. This option is only helpful in the case of small panels.

4.3 Stored results

`xtdcce2` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(N_g)`	number of groups (cross-sectional units)
`e(df_m)`	model degrees of freedom
`e(df_r)`	residual degree of freedom
`e(T)`	number of time periods
`e(K_mg)`	number of regressors (excluding variables partialed out)
`e(K_partial)`	number of partialed-out variables
`e(K_omitted)`	number of omitted variables
`e(K_pooled)`	number of pooled (homogeneous) coefficients
`e(mss)`	model sum of squares
`e(rss)`	residual sum of squares
`e(F)`	F statistic
`e(rmse)`	root mean squared error (RMSE)
`e(r2)`	R-squared
`e(r2_a)`	R-squared adjusted
`e(cd)`	CD test statistic
`e(cdp)`	p-value of CD test statistic
`e(Tmin)`	minimum time (only unbalanced panels)
`e(Tbar)`	average time (only unbalanced panels)
`e(Tmax)`	maximum time (only unbalanced panels)
`e(cr_lags)`	number of lags of cross-sectional averages

Macros

`e(cmd)`	`xtdcce2`
`e(cmdline)`	command as typed
`e(depvar)`	name of dependent variable
`e(indepvar)`	name of independent variables
`e(tvar)`	name of time variable
`e(idvar)`	name of unit variable
`e(omitted)`	omitted variables
`e(lr)`	variables in long-run cointegration vector
`e(pooled)`	pooled (homogeneous) coefficients
`e(insts)`	instruments (exogenous) variables (only IV)
`e(instd)`	instrumented (endogenous) variables (only IV)
`e(version)`	`xtdcce2` version, if `xtdcce2, version` used

Matrices

`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix of the estimators
`e(bi)`	coefficient vector of individual and pooled coefficients
`e(Vi)`	variance–covariance matrix of individual and pooled coefficients

Functions

`e(sample)`	marks estimation sample

4.4 Postestimation

`predict` and `estat` can be used after `xtdcce2`. The syntax of `predict` following `xtdcce2` is

`predict` [*type*] *newvar* [*if*] [*in*] [`, xb stdp residuals coefficients se`]

The default option is `xb`, which calculates the fitted values. `residuals` calculates the residuals, and `stdp` calculates the standard error of the prediction. `coefficients` creates a separate variable for each coefficient with the unit-specific estimate. `se` similarly creates a variable with the standard errors. The new variables have the name *newvar_varname*.

`estat` following `xtdcce2` draws a box, bar, or range plot of the MG coefficients. The syntax is

`estat` *graphtype* [*varlist*] [*if*] [*in*] [`, individual(`*string*`) combine(`*string*`) nomg cleargraph`]

graphtype is `bar` for a bar plot, `box` for a box plot, or `rcap` for a range plot. *varlist* is optional; if it is not specified, all MG coefficients are included. If the bar or range plot is drawn, then a bar plot for each MG coefficient defined by *varlist* is created, and all plots are combined in the end. The option `individual()` passes further graph options to the individual graphs, and `combine()` passes them to the combined graph. If a box plot is drawn, `individual()` controls the appearance of the graph. A confidence interval around the mean of the MG estimate is added to the range plot. Option `nomg` prevents including the confidence interval. The option `cleargraph` clears the option of the graph command and is best used in combination with the `combine()` and `individual()` options. Options `combine()` and `individual()` are used without leading and ending quotation marks. The name of the graph is saved as `r(graph_name)`.

4.5 xtcd2

Included in the `xtdcce2` package is the `xtcd2` command, which tests for weak cross-sectional dependence. The command supports balanced and unbalanced panels.[9] For a discussion of the test statistic, see section 3.

Syntax

`xtcd2` [*varname*] [*if*] [`, noestimation rho kdensity name(`*string*`)`]

varname is the name of the residuals or variables to be tested. *varname* is optional in case the command is performed after an estimation command that supports `predict, residuals`. Then, `xtcd2` predicts and tests the residuals for weak cross-sectional dependence.

9. `xtcd2` differs from existing routines such as `xtcsd` or `xtcd` in that it follows the computation of the correlation coefficients in Pesaran (2015), while other routines rely on Stata's correlation function. Therefore, a difference can occur if the average of the variable within a cross-section is nonzero.

Options

`noestimation` allows any variable to be tested. If `noestimation` is specified, then `xtcd2` is not run as a postestimation command and does not require `e(sample)` to be set. If not set, then `xtcd2` either uses the variable specified in *varname* or predicts the residuals using `predict, residuals`. In both cases, the sample is restricted to `e(sample)`.

`rho` saves the matrix with the cross correlations in `r(rho)`.

`kdensity` plots a kernel density graph of the cross correlations and reports the number of observations, the mean, percentiles, minimum, and maximum of the cross correlations. If `name(`*string*`)` is set, then the graph is saved and not drawn.

`name(`*string*`)` saves the `kdensity`.

Stored results

`xtcd2` stores the following in `r()`:

Scalars

`r(CD)`	value of the CD test statistic
`r(p)`	*p*-value of CD test statistic

Matrices

`r(rho)`	cross correlations matrix, if requested

4.6 The constant in xtdcce2

`xtdcce2` can treat the individual-specific constants α_i in several ways. In Pesaran (2006) and Chudik and Pesaran (2015a), the individual-specific constants are part of the matrix that includes the cross-sectional averages and are partialed out. In CCE regressions, the individual-specific constants include the factor loadings and the parts of the cross-sectional averages (see Chudik and Pesaran [2015b, 397]).

`xtdcce2` estimates and reports an MG estimate of the individual-specific constants if the option `reportconstant` is used. Otherwise, they are partialed out or removed from the model and not reported. Additionally, `xtdcce2` allows the constants to be the same across all units if the option `pooledconstant` is specified.[10] As a final option, the constants can be completely removed from the model by using the `noconstant` option.

The individual-specific constants are removed from the model if all parameters including the constants are constrained to be homogeneous, the cross-sectional means include all variables, and the dataset is strongly balanced. Loosely speaking, when one partials out the time averages of the dependent variable and all independent variables, the data are demeaned and a homogeneous constant is rendered to be zero. Thus, `xtdcce2` automatically removes the constant from the model to improve the estimation.

10. If `pooledconstant` is used but not `reportconstant`, the constant is internally calculated but not displayed.

If the option `reportconstant` is used, then the constant is still estimated and reported in the output.

5 Empirical examples

In this section, I carry out three empirical examples to demonstrate the use of `xtdcce2`. I fit a Solow model with DCCE and IV regression. In two other examples, I compare estimations using `xtdcce2` with the existing commands `xtmg` and `xtpmg`.

5.1 Dynamic CCE and testing for cross-sectional dependence

In the first example, I fit the Solow model in the manner of Mankiw, Romer, and Weil (1992), Islam (1995), and Lee, Pesaran, and Smith (1997). The dependent variable is the log gross domestic product (GDP) per capita, `log_rgdpo`; the independent variables are lagged GDP per capita, `L.log_rgdpo`; physical capital, `log_ck`; and the population growth rate, `log_ngd`.[11] The Penn World Tables (Feenstra, Inklaar, and Timmer 2015), version 8.0, are used and restricted to the years from 1960 to 2007, which means that there is a maximum of $T = 48$ years. Both independent variables and the level on the dependent variable are added as cross-sectional averages, set by the option `crosssectional()`. The number of cross-section averages is set to $\lfloor \sqrt[3]{48} \rfloor = 3$, specified by `cr_lags()`. Together with the first lag of `log_rgdpo` and the three lags of the cross-sectional averages, four time periods are lost, and the number of time periods used is reduced to 44. Because the cross-sectional dimension compared with the time dimension is $N/T = 93/44 = 2.11$ times larger, I apply the "half-panel" jackknife bias-correction method using the option `jackknife`.

11. In Mankiw, Romer, and Weil (1992), the dependent variable is the first difference of log GDP per capital. However, to have a lag of the dependent variable as an explanatory variable, the level is used instead. The only difference is the interpretation of the coefficient on the lagged dependent variable.

```
. use xtdcce2_sample_dataset.dta
. xtset id year
       panel variable:  id (strongly balanced)
        time variable:  year, 1960 to 2007
                delta:  1 unit
. xtdcce2 log_rgdpo L.log_rgdpo log_ck log_ngd,
> crosssectional(log_rgdpo log_ck log_ngd) cr_lags(3) jackknife
(Dynamic) Common Correlated Effects Estimator - Mean Group
Panel Variable (i): id                          Number of obs     =       4092
Time Variable (t): year                         Number of groups  =         93
                                                Obs per group (T) =         44
Degrees of freedom per group:
 without cross-sectional averages   = 41        F(1396, 2696)     =       5.10
 with cross-sectional averages      = 28        Prob > F          =       0.00
Number of                                       R-squared         =       0.73
 cross sectional lags               = 3         Adj. R-squared    =       0.58
 variables in mean group regression = 279       Root MSE          =       0.06
 variables partialled out           = 1117
                                                CD Statistic      =       0.64
                                                   p-value        =     0.5226
------------------------------------------------------------------------------
     log_rgdpo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
 Mean Group:
   L.log_rgdpo |   .6125404   .0285041    21.49   0.000      .5566735    .6684074
        log_ck |   .1151056   .0375238     3.07   0.002      .0415604    .1886508
       log_ngd |   .0451824   .1065705     0.42   0.672     -.1636919    .2540568
------------------------------------------------------------------------------
Mean Group Variables: L.log_rgdpo log_ck log_ngd
Cross Sectional Averaged Variables: log_rgdpo log_ck log_ngd
Heterogenous constant partialled out. Jackknife bias correction used.
```

On the lower right of the above results, the output shows a CD test statistic of 0.64 with a p-value of 0.52, so the null hypothesis of weak cross-sectional dependence fails to be rejected. Below the coefficient estimates, `xtdcce2` displays the names of the three MG variables and three cross-sectional averages.

A regression without any pooled variables is essentially a regression run on each country separately. The degree of freedom of a regression on each country separately is shown on the left-hand side under the time variable identifier. The first line shows the degree of freedom without the inclusion of cross-sectional averages, which results in the number of time periods used ($T = 44$) minus the number of variables ($K = 3$). In the line below, the degree of freedom for each country with cross-sectional averages is displayed. It equals the number of time periods ($T = 44$) minus the number of variables ($K = 3$) minus the number of cross-section averages times the number of lags ($p_T = 3$) plus one for the contemporaneous averages and minus one for the constant [$44-3-3\times(3+1)-1 = 28$]. In the section below, the number of lags of the cross-sectional means is displayed, together with the number of variables in the MG regression and the number of variables partialed out, which equals the number of cross-sectional averages. Because the cross-sectional averages are purely treated as controls and have no interpretation, we lose no information by partialing out. Therefore, the averages are regressed on each of the explanatory variables of interest and then the residuals collected, which are then used as the new explanatory and dependent variables. The partialing out is performed in

Mata. The variables to be partialed out (the cross-sectional means and, if requested, the heterogeneous intercept) are stacked in a block diagonal matrix, with zeros on the off diagonals. For many units, the matrix becomes sparse, and calculating and inverting the cross product becomes computationally intensive and therefore time consuming. To improve speed, the partialing out is done sequentially by unit, which is possible as long as the coefficients on the cross-sectional means, $\boldsymbol{\delta}_{il}$, are heterogeneous.[12] Within this process, the program checks whether the factor loadings are of full rank.[13] If the check fails, `xtdcce2` aborts with error code 506; see section 7. For calculating the cross-sectional averages and the partialing out, the dataset is restricted to the observations used in the regression. For the example above, this means that one period is lost to create `L.log_rgdpo`, and a further three periods are lost because of the creation of the lags of the cross-sectional means. In total, four periods are lost—one for the lag of the dependent variable and a further three for the cross-sectional averages. So the time span for the regression is the years 1964–2007, making the time dimension $T = 44$.

The regression results are in favor of the Solow model. The coefficients on physical capital and the lagged dependent variable are positive and significant, while the coefficient on population growth is positive but not significant. The estimated capital share is around 23%.[14] For a more detailed discussion of the Solow model in growth empirics, see Mankiw, Romer, and Weil (1992); Islam (1995); Lee, Pesaran, and Smith (1997); or Durlauf, Johnson, and Temple (2005); and Jones (2015). For a focus on slope heterogeneity, see Islam (1998) and Lee, Pesaran, and Smith (1998).

Use `predict, residuals` to predict the error term. The test on cross-sectional dependence can then be done by hand to confirm the result from above.

```
. predict xtdcce2_residuals, residuals
. xtcd2 xtdcce2_residuals
Pesaran (2015) test for weak cross sectional dependence
H0: errors are weakly cross sectional dependent.
        CD = 0.639
   p-value = 0.523
```

Using the option `noestimation` leads to the same result as long as the observations that are omitted in the estimation are missing in the variable `residuals`. The advantage of `noestimation` is that it does not require a sample being set by `e(sample)`; therefore, any observable variable can be tested for weak cross-sectional dependence. For example, testing the independent variable for weak cross-sectional dependence reads

12. The precision lies in a negligible order of magnitude and is offset by the improvement in speed. A simulation supporting these results is available upon request from the author. The standard solver for the calculation of the inverse of the cross product of the factor loadings is `cholsolve`. `cholsolve` cannot solve positive definite or singular matrices. In this case, use `qrsolve`. Thanks to Mark Schaffer for providing the code for this routine.
13. The condition of a full rank is checked on the unit-specific matrices containing the cross-sectional averages ($\overline{\mathbf{z}}_i = (\overline{\mathbf{y}}_i, \overline{\mathbf{x}}_i)$, where $\overline{\mathbf{y}}_i$ and $\overline{\mathbf{x}}_i$ are $T \times 1$ and $T \times K$ matrices containing the cross-sectional averages). This is possible because the matrix over all units is block diagonal, and its rank is equal to the sum of the ranks of the blocks.
14. The calculation is $\alpha = \_b[\texttt{log\_ck}]/(1-\_b[\texttt{L.log\_rgdpo}]+\_b[\texttt{log\_ck}])$. See Ditzen and Gundlach (2016) for a more detailed discussion about the estimation of the capital shares in the Solow model.

```
. xtcd2 log_rgdpo, noestimation
Pesaran (2015) test for weak cross sectional dependence
H0: errors are weakly cross sectional dependent.
        CD = 452.528
   p-value = 0.000
```

`estat` can be used for a graphical analysis of the MG regression results.

```
. estat rcap log_ngd log_ck if id <= 20
Combined graph saved as xtdcce2_combine.
```

The example above plots the MG estimates for the coefficients of `log_ngd` and `log_ck` using a range plot. The upper end of the range plot is the maximum of the 95% confidence interval, while the lower end is the minimum. The cross depicts the point estimated. Additionally, the point estimate of the MG and the 95% confidence interval are added. Here are the results.

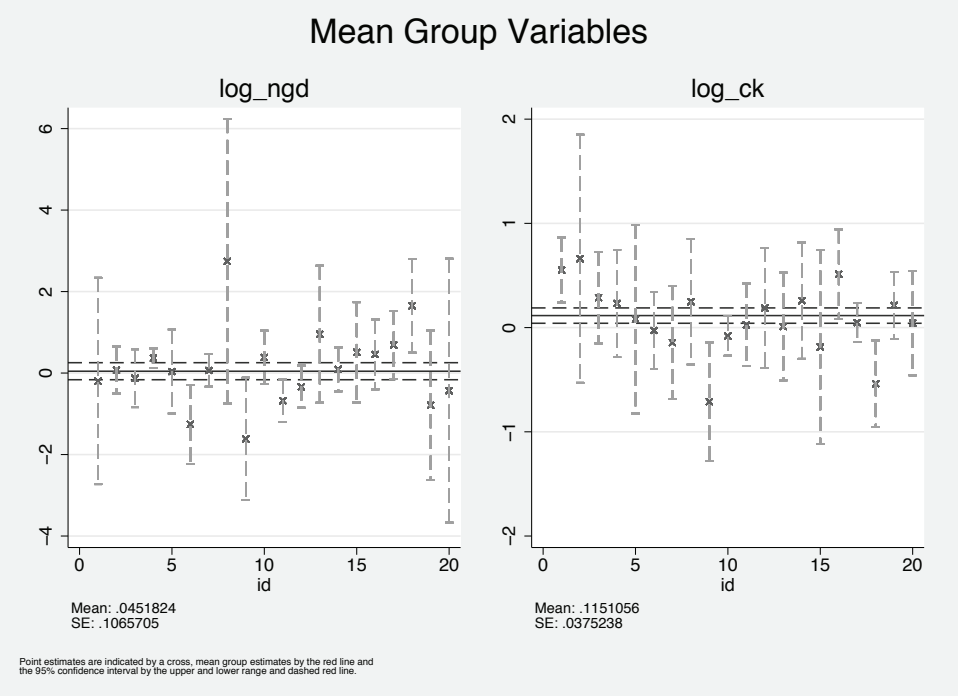

Figure 1. Range plot for MG estimates of the variables `log_ngd` and `log_ck`

Next, all three coefficients are constrained to be the same across countries ($\beta_{ik} = \beta_k, \forall i = 1, \ldots, N, k = 1, \ldots, 3$) by specifying the `pooled()` option. The constant is pooled using the `pooledconstant` option, forcing `xtdcce2` to display the constant using `reportconstant`.[15]

15. Pesaran (2006) discusses the MG and the pooled version of the CCE estimator. Chudik and Pesaran (2015b) do not mention a pooled version in the dynamic setting.

```
. xtdcce2 log_rgdpo L.log_rgdpo log_ck log_ngd,
> pooled(L.log_rgdpo log_ck log_ngd)
> crosssectional(log_rgdpo log_ck log_ngd) cr_lags(3)
> pooledconstant reportconstant
(Dynamic) Common Correlated Effects Estimator - Pooled
Panel Variable (i): id                          Number of obs      =        4092
Time Variable (t): year                         Number of groups   =          93
                                                Obs per group (T) =          44

Degrees of freedom per group:
 without cross-sectional averages   = 40        F(1120, 2972)      =        6.78
 with cross-sectional averages      = 28        Prob > F           =        0.00
Number of                                       R-squared          =        0.72
 cross sectional lags               = 3         Adj. R-squared     =        0.61
 variables in mean group regression = 4         Root MSE           =        0.06
 variables partialled out           = 1116
                                                CD Statistic       =       -0.89
                                                   p-value         =      0.3738
```

log_rgdpo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Pooled:						
L.log_rgdpo	.796726	.0661353	12.05	0.000	.6671033	.9263487
log_ck	.0847639	.0429912	1.97	0.049	.0005027	.1690251
log_ngd	.0121593	.0459543	0.26	0.791	-.0779094	.102228
_cons	2.39e-14	.9422004	0.00	1.000	-1.846679	1.846679

```
Pooled Variables:  L.log_rgdpo log_ck log_ngd _cons
Cross Sectional Averaged Variables: log_rgdpo log_ck log_ngd
```

The estimate for the constant is zero with probability 1. This result is expected, as outlined in the section above, because all coefficients are pooled, the panel is balanced, and all variables are added as cross-sectional means. For the calculation of the standard errors, `xtdcce2` follows the approach by Pesaran (2006) as outlined in section 2 and (3). Therefore, `xtdcce2` performs an MG estimation in the background to estimate $\boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_{\text{MG}}$.

As a final example, let's assume that investments into physical capital are endogenous. Countries with a large GDP per capita can save more and therefore accumulate more capital. This leads to a reversed causality of investments into physical capital and the level of GDP; for a discussion, see, for example, Durlauf, Johnson, and Temple (2005) or Temple (1999). As suggested in Temple (1999), lags of the endogenous variable are used as an instrument. To avoid a further drop in the degree of freedom by adding more variables to the model, we use the first two lags as instruments.

In line with the syntax of `ivreg2`, instrumented (endogenous) variables and the instruments are enclosed in parentheses, where the instrumented variable is followed by an equal sign and the instruments:

```
. xtdcce2 log_rgdpo L.log_rgdpo  log_ngd
> (log_ck = L.log_ck L2.log_ck),
> crosssectional(log_rgdpo log_ck log_ngd) cr_lags(3) ivreg2options(noid)
(Dynamic) Common Correlated Effects Estimator - Mean Group IV
Panel Variable (i): id                          Number of obs     =       3999
Time Variable (t): year                         Number of groups  =         93
                                                Obs per group (T) =         43
Degrees of freedom per group:
 without cross-sectional averages   = 40        F(1396, 2603)     =      23.27
 with cross-sectional averages      = 27        Prob > F          =       0.00
Number of                                       R-squared         =       0.72
 cross sectional lags               = 3         Adj. R-squared    =       0.56
 variables in mean group regression = 279       Root MSE          =       0.04
 variables partialled out           = 1117
                                                CD Statistic      =       1.12
                                                   p-value        =     0.2618
```

log_rgdpo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Mean Group:						
log_ck	.0216319	.0357938	0.60	0.546	-.0485226	.0917865
L.log_rgdpo	.5996574	.0244558	24.52	0.000	.5517248	.6475899
log_ngd	.0632321	.0905445	0.70	0.485	-.1142319	.240696

```
Mean Group Variables: L.log_rgdpo log_ngd
Cross Sectional Averaged Variables: log_rgdpo log_ck log_ngd
Endogenous Variables:  log_ck
Exogenous Variables:  L.log_ck L2.log_ck
Heterogenous constant partialled out.
```

5.2 Pooled MG

In this section, I compare `xtdcce2` with results from `xtpmg` by Blackburne and Frank (2007). `xtpmg` implements the pooled MG estimator by Pesaran, Shin, and Smith (1999) into Stata. Following Lee, Pesaran, and Smith (1997) and Pesaran, Shin, and Smith (1999), Blackburne and Frank (2007) explain the use of `xtpmg` by estimating the long-run consumption function

$$c_{it} = \theta_{0t} + \theta_{1t} y_{it} + \theta_{2t} \pi_{it} + \mu_i + \epsilon_{it}$$

where c_{it} is the log of consumption per capita, y_{it} is the log of real per capita income, and π_{it} is the inflation rate. The error correction representation is

$$\Delta c_{it} = \phi_i (c_{i,t-1} - \theta_{0i} - \theta_{1i} y_{it} - \theta_{2i} \pi_{it}) + \delta_{11i} \Delta y_{it} + \delta_{21i} \Delta \pi_{it} + \epsilon_{it} \tag{4}$$

`xtpmg` and `xtdcce2` differ in two ways: First, `xtpmg` estimates (4), while `xtdcce2` internally estimates (leaving out any cross-sectional means)

$$\Delta c_{it} = \phi_i c_{i,t-1} + \gamma_{1i} y_{it} + \gamma_{2i} \pi_{it} + \delta_{0i} + \delta_{1i} \Delta y_{it} + \delta_{2i} \Delta \pi_{it} + \epsilon_{it} \tag{5}$$

Second, `xtpmg` calculates the long-run coefficients using maximum likelihood. `xtdcce2` treats the long-run coefficients, defined in `lr()`, as further covariates and estimates (5) entirely by OLS. To calculate the long-run coefficients, I divide the coefficients by the

negative of the long-run cointegration vector to match (4), $\theta_{1i} = -\gamma_{1i}/\phi_i$. The variances are calculated using the delta method as described in the appendix. Equation (5) and the coefficients $\gamma_{1i}, \ldots, \gamma_{Ki}$ can be estimated by using `lr_options(nodivide)`.

As in Blackburne and Frank (2007), I use `jasa2.dta` to explain consumption with inflation and income after 1962.[16] The panel dataset is unbalanced and has a minimum of 31 time periods and a maximum of 32 time periods. You can assume that N/T is constant and therefore apply a CCE estimator. The output from `xtdcce2` is the following:[17]

```
. use jasa2, clear
. tsset id year
       panel variable:  id (unbalanced)
        time variable:  year, 1960 to 1993
                delta:  1 unit
. eststo xtdcce1: xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) pooled(l.c pi y) nocrosssectional lr_options(xtpmgnames)
(Dynamic) Common Correlated Effects Estimator - Pooled Mean Group
Panel Variable (i): id                          Number of obs     =        767
Time Variable (t): year                         Number of groups  =         24
                                                Obs per group:
                                                              min =         31
                                                              avg =         32
                                                              max =         32
Degrees of freedom per group:
 without cross-sectional averages   = 26.958333  F(52, 715)        =      36.62
 with cross-sectional averages      = 25.958333  Prob > F          =       0.00
Number of                                       R-squared         =       0.73
 cross sectional lags               = 0         Adj. R-squared    =       0.71
 variables in mean group regression = 51        Root MSE          =       0.02
 variables partialled out           = 1
                                                CD Statistic      =       4.10
                                                   p-value        =     0.0000
```

D.c	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Short Run Est.						
Mean Group:						
D.pi	-.0548234	.0298586	-1.84	0.066	-.1133453	.0036984
D.y	.3802491	.0350071	10.86	0.000	.3116365	.4488617
Long Run Est.						
Pooled:						
ec	-.1683577	.1195811	-1.41	0.159	-.4027324	.066017
pi	-.1941238	.1114767	-1.74	0.082	-.4126141	.0243664
y	.9025766	.1319133	6.84	0.000	.6440312	1.161122

```
Pooled Variables:  ec pi y
Mean Group Variables: D.pi D.y
Long Run Variables: ec pi y
Heterogenous constant partialled out.
```

16. The dataset is available at http://www.econ.cam.ac.uk/faculty/pesaran.
17. For later use, the regression results are stored using `eststo`.

The long-run and short-run estimates are split into two parts; one shows the results for the average long-run coefficients, and the other shows the results for the average short-run coefficients.[18] Because the dataset is unbalanced, the minimum, average, and maximum number of time periods are displayed. For the remaining regressions, `esttab` produces the following output:

```
. eststo xtpmg:  quietly xtpmg d.c d.pi d.y if year>=1962, lr(l.c pi y) ec(ec)
> replace pmg
. eststo xtdcce2: quietly xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) pooled(l.c pi y) nocrosssectional lr_options(nodivide xtpmgnames)
. eststo xtdcce3: quietly xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) pooled(l.c pi y) crosssectional(d.c d.pi d.y)
> cr_lags(0) lr_options(xtpmgnames)
. esttab xtpmg xtdcce1 xtdcce2 xtdcce3,
> mtitles("xtpmg - mg" "xtdcce2 - mg" "xtdcce2 - mg" "xtdcce2 - cce" )
> modelwidth(13) se s(N cd cdp)
```

	(1) xtpmg - mg	(2) xtdcce2 - mg	(3) xtdcce2 - mg	(4) xtdcce2 - cce
ec				
pi	-0.466***	-0.194	-0.0327	-0.276
	(0.0567)	(0.111)	(0.0473)	(0.195)
y	0.904***	0.903***	0.152**	0.940***
	(0.00868)	(0.132)	(0.0541)	(0.0895)
SR				
ec	-0.200***	-0.168	-0.168***	-0.184*
	(0.0322)	(0.120)	(0.0490)	(0.0901)
D.pi	-0.0183	-0.0548	-0.0548	0.0237
	(0.0278)	(0.0299)	(0.0299)	(0.0317)
D.y	0.327***	0.380***	0.380***	0.384***
	(0.0574)	(0.0350)	(0.0350)	(0.0431)
_cons	0.154***			
	(0.0217)			
N	767	767	767	767
cd		4.101	4.101	0.671
cdp		0.0000410	0.0000410	0.502

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Column (1) shows the results using `xtpmg`, and columns (2)–(4) show the results using the `xtdcce2` estimator. Column (1) matches the results from Blackburne and Frank (2007, 203). As expected, the MG estimates obtained by `xtpmg` and `xtdcce2` differ because of the different estimation methods. However, the signs of the MG estimates are the same, especially for the short-run coefficients. This implies that `xtdcce2` can

18. First, the long-run coefficients for each cross-section are computed. Second, the individual long-run coefficients are averaged. As an example, the average long-run coefficient for $\widehat{\theta}_1$ is calculated as $\widehat{\overline{\theta}}_1 = 1/N \sum_{i=1}^{N} \widehat{\theta}_{1,i} = 1/N \sum_{i=1}^{N} (-\widehat{\gamma}_{1,i}/\widehat{\phi}_i)$. If ϕ is heterogeneous, but let's say θ_1 is homogeneous, then the long-run coefficient γ_1 is calculated as $\gamma_1 = -\theta_1/(1/N \sum_{i=1}^{N} \phi_i)$.

be employed to estimate pooled MG estimates. In column (3), the option `nodivide` is used, producing estimates for (5).[19]

As the last row indicates, using no cross-sectional averages leads to a rejection of the null of weak cross-section dependence. Cross-sectional dependence remains in error terms, and OLS becomes inconsistent. To account for the cross-sectional dependence, I add cross-sectional averages on column (4). The p-value decreases to 0.5, and the hypothesis of weak cross-sectional dependence cannot be rejected any longer.

The average short-run coefficients can be restricted to be equal across all units by including them in the `pooled()` option. Concurrently, the average long-run coefficients can be allowed to vary as well. To test under which constraints the model is consistent, we can perform the Hausman test:

```
. eststo mg: qui xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) nocrosssectional
. eststo pmg: qui xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) pooled(l.c pi y) nocrosssectional
. eststo pooled: qui xtdcce2 d.c d.pi d.y if year >= 1962,
> lr(l.c pi y) pooled(l.c pi y d.pi d.y) nocrosssectional
. hausman mg pooled, sigmamore
```

	Coefficients			
	(b) mg	(B) pooled	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
pi				
D1.	-.0253642	-.0280826	.0027184	.
y				
D1.	.2337588	.3811944	-.1474357	.
c				
L1.	-.3063473	-.1794146	-.1269326	.
pi	-.3529095	-.266343	-.0865666	.0844914
y	.9181344	.9120574	.0060771	.

```
                b = consistent under Ho and Ha; obtained from xtdcce2
     B = inconsistent under Ha, efficient under Ho; obtained from xtdcce2
Test:  Ho:  difference in coefficients not systematic
            chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                    =        2.37
          Prob>chi2 =      0.7964
          (V_b-V_B is not positive definite)
```

19. Ditzen and Gundlach (2016) outline an alternative to obtain long-run coefficients in a dynamic panel using a restricted version of the between estimator.

```
. hausman pmg pooled, sigmamore
                ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      pmg        pooled       Difference          S.E.
-------------+----------------------------------------------------------------
          pi |
         D1. |   -.0548234    -.0280826       -.0267408             .
           y |
         D1. |    .3802491     .3811944       -.0009453             .
           c |
         L1. |   -.1683577    -.1794146        .0110569             .
          pi |   -.1941238     -.266343        .0722191        .0396237
           y |    .9025766     .9120574       -.0094807             .
------------------------------------------------------------------------------
                    b = consistent under Ho and Ha; obtained from xtdcce2
     B = inconsistent under Ha, efficient under Ho; obtained from xtdcce2

    Test:  Ho:  difference in coefficients not systematic

                  chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.97
                Prob>chi2 =      0.9650
                (V_b-V_B is not positive definite)
```

The result of the Hausman test is similar to the one obtained in Blackburn and Frank (2007, sec. 4.3 and 4.4). The first Hausman test implies that the pooled model is preferred over the MG model. The second Hausman test compares the pooled MG and the pooled model. The conclusion is different than in Blackburne and Frank (2007), that the pooled group model is preferred. However, one difference to Blackburn and Frank (2007) and one limitation of the Hausman test are worth noting. First, `xtdcce2` includes all coefficients in the Hausman test, while Blackburn and Frank (2007) include only the coefficients of the long-run vector (`pi` and `y`). Second, as Pesaran and Yamagata (2008) point out, a Hausman test lacks power in the case of pure exogenous regressors if, under the null, the slope parameters are drawn from the same distribution. Also, a test for slope homogeneity in an unknown multifactor error structure model with a lagged dependent variable in a large N and large T panel has not been established.[20]

5.3 MG and CCE

`xtdcce2` can compute the MG and CCE estimator by Pesaran and Smith (1995) and Pesaran (2006), introduced by Eberhardt (2012) via the `xtmg` command. Following Eberhardt (2012), using `manu_stata9.dta`, `xtmg` leads to the following MG results:[21]

20. Ando and Bai (2015) derive a test for a multifactor error structure model, but they assume that the common factors are estimated. An estimation of the common factors is not considered either in Pesaran (2006) or in Chudik and Pesaran (2015a) and is not supported by `xtdcce2`.
21. `manu_stata9.dta` is taken from Eberhardt and Teal (2017) and is available at https://sites.google.com/site/medevecon/.

```
. use manu_stata9.dta
. xtset nwbcode year
       panel variable:  nwbcode (strongly balanced)
        time variable:  year, 1970 to 2002
                delta:  1 unit
. eststo xtmg95: qui xtmg ly lk, trend
. eststo xtmg06: qui xtmg ly lk, cce trend
. estout xtmg95 xtmg06, c(b(star fmt(4)) se(fmt(4) par))
> mlabels("xtmg - mg" "xtmg - cce" ) s(N cd cdp, fmt(0 3 3 ))
> drop(*_ly *_lk) rename(__000007_t trend) collabels(,none)
```

	xtmg - mg	xtmg - cce
lk	0.1789*	0.3125***
	(0.0805)	(0.0849)
trend	0.0174***	0.0108**
	(0.0030)	(0.0035)
_cons	7.6528***	4.7860***
	(0.8546)	(1.3227)
N	1194	1194
cd		
cdp		

```
. eststo xtdcce95: qui xtdcce2 ly lk,
> crosssectional(ly lk) trend nocrosssectional reportconstant
. eststo xtdcce06: qui xtdcce2 ly lk,
> crosssectional(ly lk) cr_lags(0) trend reportconstant
. estout xtdcce95 xtdcce06, c(b(star fmt(4)) se(fmt(4) par))
> mlabels("xtdcce2-mg" "xtdcce2-cce" ) s(N cd cdp, fmt(0 3 3 ))
> rename(__000007_t trend) collabels(,none)
```

	xtdcce2-mg	xtdcce2-cce
lk	0.1789*	0.3125***
	(0.0805)	(0.0849)
trend	0.0174***	0.0108**
	(0.0030)	(0.0035)
_cons	7.6354***	4.7752***
	(0.8531)	(1.3202)
N	1194	1194
cd	6.686	-0.201
cdp	0.000	0.841

The first table shows the estimation results from table 1 in Eberhardt (2012, 67). The lower table displays results on the same equation using `xtdcce2`. The first column shows an MG regression in both tables, and the second column shows a CCE regression with contemporaneous cross-sectional means. The CD test statistic rejects the hypothesis of weak cross-sectional dependence in the case of the MG regression. Including cross-sectional averages improves the statistic such that the hypothesis cannot be rejected any longer. Estimation results produced by `xtmg` and `xtdcce2` differ slightly, as seen

here by the constant, because `xtdcce2` ensures that all variables are stored as `doubles` to allow for the most precision.[22]

6 Monte Carlo simulation

In this section, I perform a Monte Carlo simulation to shed light on the size of the bias of the DCCE estimator and give guidance under which dimensions the DCCE estimator can be used. I carry out the Monte Carlo simulation along the lines of Chudik and Pesaran (2015b). The underlying model is

$$
\begin{aligned}
y_{it} &= c_{yi} + \phi_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + u_{it} \\
u_{it} &= \gamma_i f_t + \epsilon_{it} \\
x_{it} &= c_{xi} + \alpha_{xi} y_{i,t-1} + \gamma_{xi} f_t + v_{xit} \\
g_{it} &= c_{gi} + \alpha_{gi} y_{i,t-1} + \gamma_{gi} f_t + v_{git}
\end{aligned}
$$

where y_{it} is the dependent variable and x_{it} a vector of K independent variables. Without loss of generality, it is assumed that only one independent variable exists. g_{it} is a set of covariates that are affected by the unobserved factors but are not used to estimate y_{it}. The coefficient for the contemporaneous value of x_{it} is drawn from a uniform distribution as $\beta_{0i} \sim \text{IIDU}(0.5, 1)$. The coefficient on the lagged value of the independent variable is set to $\beta_{1i} = -0.5$. ϕ_i and α_{xi} depend on each other to make sure that the series y_{it} and x_{it} are stationary.[23] They are specified as $\phi_i \sim \text{IIDU}(0, 0.8)$ and $\alpha_{xi} \sim \text{IIDU}(0, 0.35)$. The equation for g_{it} is independent, and α_{gi} is set to $\alpha_{gi} \sim \text{IIDU}(0, 1)$. Compared with Chudik and Pesaran (2015b), the number of common factors is restricted to one. As shown in their Monte Carlo simulation, the results are robust for a few common factors. The common factors f_t are potentially correlated over time ($\rho_f \neq 0$) and calculated as $f_t = \rho_f f_{t-1} + \varsigma_{ft}, \varsigma_{ft} \sim \text{IIDN}(0, 1 - \rho_f^2)$. The error ϵ_{it} is heteroskedastic and weakly cross-sectionally dependent ($\alpha_{\text{CSD}} = 0.4$). Appendix A.2 describes the data-generating process (DGP) in more detail.

The DCCE estimator is exposed to three potential sources for a bias: the length of the time series, cross-sectional dependence, and heterogeneous slope coefficients. The first source relates to small T and the time-series bias of order T^{-1} (Hurwicz bias) and is expected to decrease with $T \to \infty$. The bias due to the heterogeneous slope coefficients is expected to decrease with N because the MG coefficients are calculated over a larger number of cross-sectional units. The bias due to the cross-sectional dependence should decrease with both N and T. The direction of the bias will become apparent by comparing results with and without cross-sectional averages.

22. Another difference from `xtmg` is that `xtdcce2` supports time-series operators.
23. See Chudik and Pesaran (2015b, 399–400).

The number of cross-sectional unit and time periods varies from 40 to 200. I present results with and without cross-sectional averages to show the influence of the cross-sectional averages. `xtdcce2` estimates the following equation:

$$y_{it} = c_{yi} + \phi_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + \sum_{l=0}^{p_t} \boldsymbol{\delta}_{il}' \overline{\mathbf{z}}_{t-l} + e_{yit}$$

The number of lags is set to the integer part of $T^{\frac{1}{3}}$. The cross-sectional averages contain y, x, and g. Within each run of the Monte Carlo simulation, the following command line for `xtdcce2` was used:

```
. xtdcce2 y L.y x L.x , cr_lags(lags) crosssectional(y x)
```

Table 1 presents the Monte Carlo simulation results with contemporaneous cross-sectional averages and three lags. The first panel shows the bias and root mean squared error (RMSE) for the MG estimates of the autocorrelation coefficient ϕ, the second panel shows the MG estimates of the coefficient on x, and the last panel shows the MG estimates of the coefficient on the lagged value of x. The bias of ϕ is substantial with a small number of time periods and cross-sectional units. However, the bias decreases with an increase in time periods but remains almost constant if the number of cross-sectional units is increased. Similarly to the bias, the RMSE decreases with an increase in the number of time periods, implying a decrease in the variation of the bias as well.

Table 1. Monte Carlo results for ϕ, β_0, and β_1, with $p_T = \left[T^{1/3}\right]$. The DGP is $y_{it} = c_{yi} + \phi_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + \gamma_i f_t + \epsilon_{it}$, where $\phi = E(\phi_i) = 0.4$, $\beta_{0i} \sim \text{IIDU}(0.5, 0.1)$, $\beta_{1i} = -0.5$, $c_{yi} \sim \text{IIDN}(0,1)$, $\gamma_i = \sqrt{1-\sigma_\gamma^2} + \eta_{i\gamma}$ with $\eta_{i\gamma} \sim \text{IIDN}(0, \sigma_\gamma^2)$ and $\sigma_\gamma^2 = 0.2^2$, $f_t = \rho_f f_{t-1} + \varsigma_{ft}$ with $\varsigma_{ft} \sim \text{IIDN}(0, 1-\rho_f^2)$, $\rho_f = 0.6$. $\alpha_{\text{CSD}} = 0.4$. For a further description, see section 6.

(N,T)	Bias (×100)					RMSE (×100)				
	40	50	100	150	200	40	50	100	150	200
$\phi = 1/N \sum_{i=1}^{N} \phi_i$										
40	−42.85	−31.69	−13.52	−8.06	−5.54	18.14	13.53	6.26	4.11	3.02
50	−42.91	−30.28	−13.45	−8.25	−6.33	18.03	12.95	6.11	3.99	3.18
100	−43.33	−31.51	−13.66	−8.83	−6.17	17.86	12.96	5.85	3.90	2.78
150	−42.16	−31.11	−13.73	−8.74	−6.31	17.23	12.70	5.70	3.72	2.75
200	−43.65	−31.43	−13.69	−8.96	−6.19	17.75	12.80	5.67	3.73	2.65
$\beta_0 = 1/N \sum_{i=1}^{N} \beta_{0i}$										
40	4.34	3.37	2.05	1.30	0.67	12.47	9.21	5.62	4.40	3.66
50	4.78	3.10	2.13	1.26	1.24	11.10	8.65	5.32	4.25	3.77
100	4.27	3.71	2.17	1.15	1.18	8.54	6.55	4.10	3.11	2.62
150	3.66	3.96	2.07	1.30	1.02	7.04	5.94	3.47	2.56	2.15
200	5.21	4.07	2.34	1.39	0.96	6.71	5.17	3.18	2.37	1.89
$\beta_1 = 1/N \sum_{i=1}^{N} \beta_{1i}$										
40	−8.61	−6.20	−1.68	−0.91	−0.91	12.07	10.20	6.02	4.31	3.97
50	−6.82	−5.27	−1.83	−1.02	−1.24	11.54	9.25	5.55	4.06	3.67
100	−5.12	−3.04	−1.07	−1.10	−0.16	8.14	6.41	3.75	3.15	2.66
150	−6.78	−2.76	−0.45	−0.39	−0.29	7.50	5.88	3.32	2.65	2.15
200	−5.13	−2.88	−0.44	−0.68	−0.21	6.63	5.07	2.79	2.31	1.85

The results for the two other coefficients are similar. The bias decreases with further time periods but remains at a similar level with an increase in the number of the cross-sectional units. Note that the MG coefficients identify the homogeneous coefficient β_1 well. Thus, one can conclude that the MG estimator can be applied to a coefficient for which the heterogeneity assumption is questioned at little costs.

For all three coefficients, the decrease in the bias with $T \to \infty$ implies that the small-sample time-series bias plays a substantial role. The bias due to the heterogeneous slope coefficients appears to be less important.

To assess the bias due to cross-sectional dependence, I run the estimations without cross-sectional averages using the option `nocrosssectional`:

```
. xtdcce2 y L.y x L.x, nocrosssectional crosssectional(y x)
```

Compared with table 1, the bias of ϕ for low values of T appears to be slightly smaller, but it does not shrink as much with $T \to \infty$ as in the case with cross-sectional averages. The bias with cross-sectional averages reduces to -6.19%, while without the averages

the bias is −15.17%. The difference can be interpreted as the bias because it does not account for cross-sectional dependence. The bias for β_0 is larger by a magnitude and up to 80.86%. The results for the homogeneous coefficient β_1 are noteworthy because the bias changes from a negative into a positive area with an increase in T.

Table 2. Monte Carlo Results for ϕ, β_0, and β_1, and no cross-sectional averages. See table 1 for notes.

(N,T)	Bias (×100)					RMSE (×100)				
	40	50	100	150	200	40	50	100	150	200
$\phi = 1/N \sum_{i=1}^{N} \phi_i$										
40	−30.37	−25.79	−17.59	−14.95	−13.76	12.82	10.90	7.55	6.43	5.96
50	−31.32	−26.26	−17.76	−15.49	−14.27	13.05	11.07	7.59	6.55	6.05
100	−31.72	−27.04	−18.88	−16.12	−14.79	12.99	11.05	7.77	6.65	6.06
150	−31.36	−26.65	−18.90	−16.35	−15.08	12.75	10.85	7.68	6.66	6.15
200	−31.54	−27.20	−18.81	−16.61	−15.17	12.76	11.01	7.65	6.71	6.15
$\beta_0 = 1/N \sum_{i=1}^{N} \beta_{0i}$										
40	79.72	80.19	77.55	77.69	77.62	60.82	60.86	58.60	58.47	58.31
50	80.24	78.49	78.11	77.97	77.45	61.11	59.61	58.95	58.62	58.30
100	80.57	79.90	78.51	77.63	77.84	61.10	60.42	59.05	58.39	58.49
150	80.12	79.14	77.66	77.81	77.66	60.58	59.82	58.42	58.49	58.38
200	80.86	79.84	78.24	77.72	77.27	61.06	60.22	58.90	58.37	58.03
$\beta_1 = 1/N \sum_{i=1}^{N} \beta_{1i}$										
40	−3.51	0.10	5.03	6.89	8.65	10.01	8.84	6.37	6.34	6.18
50	−4.41	−1.15	5.87	7.46	7.86	9.10	7.99	6.25	6.01	5.69
100	−4.26	0.18	5.39	6.55	7.49	7.02	5.79	4.87	4.83	4.89
150	−4.21	−0.87	4.81	6.36	7.40	6.31	5.30	4.29	4.43	4.58
200	−3.10	−0.20	4.40	6.37	7.37	6.03	4.69	3.82	4.16	4.32

The Monte Carlo simulation emphasizes the importance of including cross-sectional averages and shows that the estimator is less biased for a large number of cross-sectional units and time periods. In the case of omitting the cross-sectional averages, MG estimates are upward biased.

7 Error messages

`xtdcce2` produces the following error codes:

`r(109)`	`ivreg2 not installed.`
`r(184)`	options `noconstant` and `pooledconstant`, `trend` and `trendconstant` or `jackknife` and `recursive` are combined.
`r(506)`	`Rank condition on cross section means not satisfied.`
`r(2001)`	`More variables than observations.`

8 Conclusion

The community-contributed command `xtdcce2` introduces new routines to fit a heterogeneous panel model using dynamic CCE in large N and T panels. It combines estimation procedures proposed in Pesaran and Smith (1995) and Pesaran, Shin, and Smith (1999) with those in Pesaran (2006) and Chudik and Pesaran (2015b). It allows coefficients to be pooled or estimated as MGs. Furthermore, it supports unbalanced panels, estimation of IVs and small-sample time-series bias corrections and tests for cross-sectional dependence using the included `xtcd2` routine. I gave an empirical example estimating a growth regression and carried out a simulation exercise to prove the estimators' consistency.

9 Acknowledgments

I am grateful to Arnab Bhattacharjee, David M. Drukker, Markus Eberhardt, Erich Gundlach and Mark Schaffer, two anonymous referees, and the participants of the 2016 Stata Users Group meeting in London for many valuable comments and suggestions. I am grateful to Achim Ahrens for his contributions to the `xtcd2` command and to Kyle McNabb for testing the command. The code and especially the output greatly benefited from Markus Eberhardt's `xtmg` command. Any remaining errors are my own.

10 References

Ando, T., and J. Bai. 2015. A simple new test for slope homogeneity in panel data models with interactive effects. *Economics Letters* 136: 112–117.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3: 1–31.

———. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 7: 465–506.

Blackburne, E. F., III, and M. W. Frank. 2007. Estimation of nonstationary heterogeneous panels. *Stata Journal* 7: 197–208.

Bond, S. R., and M. Eberhardt. 2013. Accounting for unobserved heterogeneity in panel time series models. Mimeo: University of Oxford.

Chudik, A., and M. H. Pesaran. 2015a. Large panel data models with cross-sectional dependence: A survey. In *The Oxford Handbook Of Panel Data*, ed. B. H. Baltagi, 2–45. Oxford: Oxford University Press.

———. 2015b. Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics* 188: 393–420.

Chudik, A., M. H. Pesaran, and E. Tosetti. 2011. Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal* 14: C45–C90.

De Hoyos, R. E., and V. Sarafidis. 2006. Testing for cross-sectional dependence in panel-data models. *Stata Journal* 6: 482–496.

Ditzen, J., and E. Gundlach. 2016. A Monte Carlo study of the BE estimator for growth regressions. *Empirical Economics* 51: 31–55.

Durlauf, S. N., P. A. Johnson, and J. R. W. Temple. 2005. Growth econometrics. In *Handbook of Economic Growth, vol. 1A*, ed. P. Aghion and S. N. Durlauf, 555–677. Amsterdam: North-Holland.

Eberhardt, M. 2011. xtcd: Stata module to investigate variable/residual cross-section dependence. Statistical Software Components S457237, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457237.html.

———. 2012. Estimating panel time-series models with heterogeneous slopes. *Stata Journal* 12: 61–71.

Eberhardt, M., C. Helmers, and H. Strauss. 2013. Do spillovers matter when estimating private returns to R&D? *Review of Economics and Statistics* 95: 436–448.

Eberhardt, M., and F. Teal. 2017. The magnitude of the task ahead: Productivity analysis with heterogeneous technology. Discussion Paper GEP 17/16, University of Nottingham, Globalisation and Economic Policy. https://www.nottingham.ac.uk/gep/documents/papers/2017/2017-16.pdf.

Everaert, G., and I. De Vos. 2016. Bias-corrected common correlated effects pooled estimation in homogeneous dynamic panels. Working Paper 16/920, Ghent University, Faculty of Economics and Business Administration.

Everaert, G., and T. D. Groote. 2016. Common correlated effects estimation of dynamic panels with cross-sectional dependence. *Econometric Reviews* 35: 428–463.

Feenstra, R. C., R. Inklaar, and M. P. Timmer. 2015. The next generation of the Penn World Table. *American Economic Review* 105: 3150–3182.

Gundlach, E., and M. Paldam. 2016. Socioeconomic transitions as common dynamic processes. Economics Working Papers 2016-6, Department of Economics and Business Economics, Aarhus University. https://ideas.repec.org/p/aah/aarhec/2016-06.html.

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Islam, N. 1995. Growth empirics: A panel data approach. *Quarterly Journal of Economics* 110: 1127–1170.

———. 1998. Growth empirics: A panel data approach—A reply. *Quarterly Journal of Economics* 113: 325–329.

Jann, B. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s455001.html.

Jones, C. I. 2015. The facts of economic growth. NBER Working Paper No. 21142, The National Bureau of Economic Research. http://www.nber.org/papers/w21142.

Kapetanios, G., M. H. Pesaran, and T. Yamagata. 2011. Panels with non-stationary multifactor error structures. *Journal of Econometrics* 160: 326–348.

Lee, K., M. H. Pesaran, and R. Smith. 1997. Growth and convergence in a multi-country empirical stochastic Solow model. *Journal of Applied Econometrics* 12: 357–392.

———. 1998. Growth empirics: A panel data approach—A comment. *Quarterly Journal of Economics* 113: 319–323.

Mankiw, N. G., D. Romer, and D. N. Weil. 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437.

McNabb, K., and P. LeMay-Boucher. 2014. Tax structures, economic growth and development. ICTD Working Paper 22, International Centre for Tax and Development. https://ssrn.com/abstract=2496470.

Pesaran, M. H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967–1012.

———. 2015. Testing weak cross-sectional dependence in large panels. *Econometric Reviews* 34: 1089–1117.

Pesaran, M. H., Y. Shin, and R. P. Smith. 1999. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association* 94: 621–634.

Pesaran, M. H., and R. P. Smith. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68: 79–113.

Pesaran, M. H., and T. Yamagata. 2008. Testing slope homogeneity in large panels. *Journal of Econometrics* 142: 50–93.

Temple, J. 1999. The new growth evidence. *Journal of Economic Literature* 37: 112–156.

About the author

Jan Ditzen is a research associate at the Centre for Energy Economics Research and Policy at Heriot-Watt University in Edinburgh, UK. His research interests are in the field of applied econometrics, with a focus on growth empirics and spatial econometrics, particularly cross-sectional dependence in large panels.

A Technical appendix

A.1 Delta method

For calculating the long-run coefficients, divide the estimates of $\widehat{\gamma}_{k,i}$ obtained by OLS from (5) by estimates of the error-correction speed of adjustment parameter cointegration vector $\widehat{\phi}_i$. To calculate the variance–covariance matrix, use the delta method, which

allows the calculation of an approximate probability distribution for a matrix function $\mathbf{a}(\boldsymbol{\beta})$ based on a random vector with a known variance (see, for example, Hayashi [2000, 93]). Suppose that, for the random vector, $\boldsymbol{\beta}_i \to_p \boldsymbol{\beta}$ and $\sqrt{n}(\boldsymbol{\beta}_i - \boldsymbol{\beta}) \to_d N(\mathbf{0}, \boldsymbol{\Sigma})$. Denote the first derivatives of $\mathbf{a}(\boldsymbol{\beta})$ as

$$\mathbf{A}(\boldsymbol{\beta}) \equiv \frac{\partial \mathbf{a}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$$

Then the distribution of the function $\mathbf{a}(\cdot)$ is

$$\sqrt{n}\left\{\mathbf{a}(\boldsymbol{\beta}_i) - \mathbf{a}(\boldsymbol{\beta})\right\} \to_d N\left\{\mathbf{0}, \mathbf{A}(\boldsymbol{\beta})\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\beta})'\right\}$$

For calculating the long-run coefficients and using the notation from (5), assume that

$$\boldsymbol{\beta}_i = (\phi_i, \gamma_{1,i}, \gamma_{2,i}, \delta_{1,i}, \delta_{2,i})'$$

The variance–covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} V(\phi_i) & \text{Cov}(\phi_i\gamma_{1,i}) & \text{Cov}(\phi_i\gamma_{2,i}) & \text{Cov}(\phi_i\delta_{1,i}) & \text{Cov}(\phi_i\delta_{2,i}) \\ \text{Cov}(\phi_i\gamma_{1,i}) & V(\gamma_{1,i}) & \text{Cov}(\gamma_{1,i}\gamma_{2,i}) & \dots & . \\ . & & . & & . \\ . & & & . & . \\ \text{Cov}(\phi_i\delta_{2,i}) & \dots & \dots & \dots & V(\delta_{2,1}) \end{pmatrix}$$

The function $\mathbf{a}(\cdot)$ maps the long-run coefficients and leaves the short-run coefficients:

$$\begin{aligned} \mathbf{a}(\boldsymbol{\beta}_i) &= (\phi_i, -\gamma_{1,i}/\phi_i, -\gamma_{2,i}/\phi_i, \delta_{1,i}, \delta_{2,i})' \\ &= (\phi_i, \theta_{1,i}, \theta_{2,i}, \delta_{1,i}, \delta_{2,i})' \end{aligned}$$

The first derivative of $\mathbf{a}(\cdot)$ is then

$$\mathbf{A}_i(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial\phi_i}{\partial\phi_i} & \frac{\partial\phi_i}{\partial\gamma_{1,i}} & \frac{\partial\phi_i}{\partial\gamma_{2,i}} & \frac{\partial\phi_i}{\partial\delta_{1,i}} & \frac{\partial\phi_i}{\partial\delta_{2,i}} \\ \frac{\partial\theta_{1,i}}{\partial\phi_i} & \frac{\partial\theta_{1,i}}{\partial\gamma_{1,i}} & \frac{\partial\theta_{1,i}}{\partial\gamma_{2,i}} & \frac{\partial\theta_{1,i}}{\partial\delta_{1,i}} & \frac{\partial\theta_{1,i}}{\partial\delta_{2,i}} \\ \frac{\partial\theta_{2,i}}{\partial\phi_i} & \frac{\partial\theta_{2,i}}{\partial\gamma_{1,i}} & \frac{\partial\theta_{2,i}}{\partial\gamma_{2,i}} & \frac{\partial\theta_{2,i}}{\partial\delta_{1,i}} & \frac{\partial\theta_{2,i}}{\partial\delta_{2,i}} \\ \frac{\partial\delta_{1,i}}{\partial\phi_i} & \frac{\partial\delta_{1,i}}{\partial\gamma_{1,i}} & \frac{\partial\delta_{1,i}}{\partial\gamma_{2,i}} & \frac{\partial\delta_{1,i}}{\partial\delta_{1,i}} & \frac{\partial\delta_{1,i}}{\partial\delta_{2,i}} \\ \frac{\partial\delta_{2,i}}{\partial\phi_i} & \frac{\partial\delta_{2,i}}{\partial\gamma_{1,i}} & \frac{\partial\delta_{2,i}}{\partial\gamma_{2,i}} & \frac{\partial\delta_{2,i}}{\partial\delta_{1,i}} & \frac{\partial\delta_{2,i}}{\partial\delta_{2,i}} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{\gamma_{1,i}}{\phi_i^2} & -\frac{1}{\phi_i} & 0 & 0 & 0 \\ \frac{\gamma_{2,i}}{\phi_i^2} & 0 & -\frac{1}{\phi_i} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

All components of the variance–covariance matrix are then known, and it can be calculated as

$$\boldsymbol{\Sigma}_a = \mathbf{A}(\boldsymbol{\beta})\boldsymbol{\Sigma}\mathbf{A}(\boldsymbol{\beta})'$$

A.2 Monte Carlo setup

As in Chudik and Pesaran (2015b), the DGPs are the following:

$$
\begin{aligned}
y_{it} &= c_{yi} + \phi_i y_{i,t-1} + \beta_{0i} x_{it} + \beta_{1i} x_{i,t-1} + u_{it} \\
u_{it} &= \gamma_i f_t + \epsilon_{it} \\
x_{it} &= c_{xi} + \alpha_{xi} y_{i,t-1} + \gamma_{xi} f_t + v_{xit} \\
g_{it} &= c_{gi} + \alpha_{gi} y_{i,t-1} + \gamma_{gi} f_t + v_{git}
\end{aligned}
$$

y_{it} is the dependent variable and x_{it} the only independent variable. For a matter of ease, it is assumed that only one explanatory variable exists. g_{it} is another independent variable that is affected by the unobserved factors and y_{it} but is not used to estimate it.

Common factors. The common factors are calculated as

$$
\begin{aligned}
f_t &= \rho_f f_{t-1} + \varsigma_{ft}, \quad \varsigma_{ft} \sim \text{IIDN}(0, 1-\rho_f^2) \\
v_{xit} &= \rho_{xi} v_{xi,t-1} + \varsigma_{xit}, \quad \varsigma_{xit} \sim \text{IIDN}(0, \sigma_{vxi}^2) \\
v_{git} &= \rho_{gi} v_{gi,t-1} + \varsigma_{git}, \quad \varsigma_{git} \sim \text{IIDN}(0, \sigma_{vgi}^2) \\
\rho_{xi} &\sim \text{IIDU}(0, 0.95) \\
\rho_{gi} &\sim \text{IIDU}(0, 0.95) \\
\rho_f &= 0 \text{ if serially uncorrelated factors, or if correlated } \rho_f = 0.6 \\
\sigma_{vxi}^2 = \sigma_{vgi}^2 &= \sigma_{vi}^2 = \left(\beta_{0i} \sqrt{1 - \{E(\rho_{xi})\}^2} \right)^2
\end{aligned}
$$

Fixed effects. The cross-section–specific fixed effects are generated as

$$
\begin{aligned}
c_{yi} &\sim \text{IIDN}(1, 1) \\
c_{xi} &= c_{yi} + \varsigma_{c_x i}, \quad \varsigma_{c_x i} \sim \text{IIDN}(0, 1) \\
c_{gi} &= c_{yi} + \varsigma_{c_g i}, \quad \varsigma_{c_g i} \sim \text{IIDN}(0, 1)
\end{aligned}
$$

Dependence between x_{it}, g_{it} and c_{yi} is introduced by adding c_{yi} to the equations for c_{xi} and c_{gi}.

Coefficients. The coefficient for the contemporaneous value of x_{it} is drawn from a uniform distribution as $\beta_{0i} \sim \text{IIDU}(0.5, 1)$. The coefficient on the lagged value of the independent variable is set to $\beta_{1i} = -0.5$. For the lagged dependent variable, two different scenarios are considered for calculating y_{it} and x_{it}: one with low values for ϕ, $\phi_i \sim \text{IIDU}(0, 0.8)$, and $\alpha_{xi} \sim \text{IIDU}(0, 0.35)$; and one with high values for $\phi_i \sim \text{IIDU}(0.5, 0.9)$ and $\alpha_{xi} \sim \text{IIDU}(0, 0.15)$. α_{gi} is in both scenarios the same: $\alpha_{gi} \sim \text{IIDU}(0, 1)$.[24]

24. ϕ_i and α_{xi} depend on each other to make sure that the series y_{it} and x_{it} are stationary. See Chudik and Pesaran (2015b, 399–400).

Factor loadings.

$$
\begin{aligned}
\gamma_i &= \gamma + \eta_{i\gamma}, \quad \eta_{i\gamma} \sim \text{IIDN}(0, \sigma^2_\gamma) \\
\gamma_{xi} &= \gamma_x + \eta_{i\gamma x}, \quad \eta_{i\gamma x} \sim \text{IIDN}(0, \sigma^2_{\gamma x}) \\
\gamma_{gi} &= \gamma_g + \eta_{i\gamma g}, \quad \eta_{i\gamma g} \sim \text{IIDN}(0, \sigma^2_{\gamma g}) \\
\sigma^2_\gamma &= \sigma^2_{\gamma x} = \sigma^2_{\gamma g} = 0.2^2 \\
\gamma &= \sqrt{b_\gamma}, \quad b_\gamma = \frac{1}{m} - \sigma^2_\gamma \\
\gamma_x &= \sqrt{b_x}, \quad b_x = \frac{2}{m(m+1)} - \frac{2}{m+1}\sigma^2_{\gamma x} \\
\gamma_g &= \sqrt{b_g}, \quad b_g = \frac{1}{m^2} - \frac{\sigma^2_g}{m}
\end{aligned}
$$

where m is the number of unobserved factors. Compared with Chudik and Pesaran (2015b), it is restricted to 1.

Error term. The errors are generated such that heteroskedasticity and weakly cross-sectional dependence is allowed.

$$
\begin{aligned}
\boldsymbol{\epsilon}_t &= \alpha_{\text{CSD}} \mathbf{S}_\epsilon \boldsymbol{\epsilon}_t + \mathbf{e}_{\boldsymbol{\epsilon t}} \\
\Rightarrow \boldsymbol{\epsilon}_t &= \left(\mathbf{I} - \alpha_{\text{CSD}} \mathbf{S}_\epsilon\right)^{-1} \mathbf{e}_{\boldsymbol{\epsilon t}} \\
\mathbf{e}_{\boldsymbol{\epsilon t}} &\sim \text{IIDN}\left(0, \frac{1}{2}\sigma^2_i\right), \quad \text{with } \sigma^2_i \sim \chi^2(2) \\
\mathbf{S}_\epsilon &= \begin{pmatrix}
0 & \frac{1}{2} & 0 & 0 & \dots & 0 \\
\frac{1}{2} & 0 & 1 & 0 & & 0 \\
0 & 1 & 0 & \ddots & & \vdots \\
0 & 0 & \ddots & \ddots & 1 & 0 \\
\vdots & & & 1 & 0 & \frac{1}{2} \\
0 & 0 & \dots & 0 & \frac{1}{2} & 0
\end{pmatrix}
\end{aligned}
$$