Penalized Indirect Inference*

Francisco Blasques[a,b] and Artem Duplinskiy[a]

(a) VU University Amsterdam

(b) Tinbergen Institute

January 27, 2016

Abstract

Parameter estimates of structural economic models are often difficult to interpret at the light of the underlying economic theory. Bayesian methods have become increasingly popular as a tool for conducting inference on structural models since priors offer a way to exert control over the estimation results. Similarly to Bayesian estimation, this paper proposes a penalized indirect inference estimator that allows researchers to obtain economically meaningful parameter estimates in a frequentist setting. The asymptotic properties of the estimator are established for both correctly and incorrectly specified models, as well as under strong and weak parameter identification. A Monte Carlo study reveals the role of the penalty function in shaping the finite sample distribution of the estimator. The advantages of using this estimator are highlighted in the empirical study of a state-of-the-art dynamic stochastic general equilibrium model.

JEL Codes: C15; C13; D58; E32.
Keywords: Penalized estimation, Indirect Inference, Simulation-based methods, DSGE.

---

*The authors are thankful to Ronald Gallant, Siem Jan Koopman, Andre Lucas, Tony Smith, and the participants of the *Conference on Indirect Estimation Methods in Finance and Economics*, in Konstanz, Germany, 2014, and *Nordic Econometric Meeting*, in Helsinki, 2015, and the participants of the *VU Econometrics Seminar*, Amsterdam, 2014, for helpful comments and discussion. Corresponding address: VU University Amsterdam, FEWEB/EOR, de Boelelaan 1105, 1081 HV Amsterdam, Netherlands, email: f.blasques@vu.nl.

1 Introduction

In economics, it is often difficult to reconcile the estimates obtained for parameters of structural models with the underlying economic theory. This problem is especially evident when employing frequentist estimation techniques that leave the researcher unable to exert some control over the estimator, at least within the parameter space bounds. Such problems are well known in the dynamic stochastic general equilibrium (DSGE) literature. Canova and Sala (2009) show that identification problems are pervasive in New Keynesian DSGE models; see also Ma (2002), Beyer and Farmer (2004), Nason and Smith (2008). Model misspecification also leads to difficulties since parameter estimates are biased and multiple pseudo-true parameters may exist.

In the case of DSGE models, the problem of model misspecification and the challenge posed by parameters that are unidentified, or only weakly identified, has lead to the widespread adoption of Bayesian methods. The introduction of priors in Bayesian estimation allows the researcher to exert control over the estimation procedure, in the effort to obtain economically meaningful parameter estimates, by including information that is not contained in the data sample.

For example, An and Schorfheide (2007) point out that *"Any estimation and evaluation method is confronted with the following challenges: potential model misspecification and possible lack of identification of parameters of interest."* and that Bayesian methods become useful since *"prior distributions can be used to incorporate additional information into the parameter estimation."*. Similarly, Fernandez-Villaverde (2010) defends the use of additional structure in estimation since *"Pre-sample information is often amazingly rich and considerably useful and not taking advantage of it is an unforgivable omission"*. He further concludes that *"Yes, our inference would have depended heavily on the prior, but why is this situation any worse than not being able to say anything of consequence?"*.

This paper proposes a tool for incorporating pre-sample information in an indirect inference estimation setting. The added structure is provided by a penalty function that plays a role that is similar to that played by the prior in Bayesian estimation. The penalty function is allowed to be data dependent. This is similar in spirit to the empirical Bayes method; see e.g. Morris (1983).

Penalized estimation is not new in the frequentist literature. On the contrary, the basic idea of adding penalties to estimation criterion functions such as least squares, maximum likelihood or the method of moments is present in many statistical applications.[1] This paper 'brings' the use of penalties to a simulation-based estimation setting, which allows us to estimate complex, high-dimensional, nonlinear dynamic models with unobserved variables, like DSGE models. In particular, we introduce a penalty in the criterion function of the indirect inference (II) estimator proposed by Gourieroux et al. (1993) and Smith (1993). The penalized indirect inference (PenII) estimator possesses a number of features, similar to those of Bayesian methods, that may be valuable for the estimation of complex structural economic models such as DSGE models.

First, the PenII estimator allows for a wide range of criterion functions. Specifically, the PenII estimator encompasses both limited information estimation methods, such as the simulated method of moments of Duffie and Singleton (1990) and Lee and Ingram (1991), as well as likelihood-based full information estimation methods, such as the efficient method of Gallant and Tauchen (1996). Popular indirect inference criteria include moment matching, matching VAR parameters, and matching impulse response functions at certain periods, or maximizing the likelihood of the selected auxiliary statistics; see e.g. Christiano et al. (2005), Ruge-Murcia (2007), and Dupor et al. (2009), and Creel and Kristensen (2013) for examples of these different approaches. PenII is similar to Bayesian estimation in this respect as the latter can also be performed with different criteria; see e.g. Gallant and Tauchen (2015) for Bayesian estimation with GMM type objective functions.

The ability to choose among different criteria is important for estimation of potentially misspecified models. Indeed, Fukac and Pagan (2010) argue for the use of limited information estimation techniques in DSGE models since the maximum likelihood (ML) estimator requires the entire probabilistic structure of the model to be well specified, rather than just a few features of interest. Furthermore, ML estimators may have poor robustness properties; see e.g. the seminal work of Huber (1967, 1974).

Second, the penalty function can take a wide range of forms. It is thus easy to incorporate various forms of pre-sample information in a flexible way. Strong parameter restrictions can be imposed by letting the penalty diverge to infinity. Areas of indifference in the parameter space can characterized by plateaus in the penalty

[1] Examples include the lasso penalties used e.g. in Zou (2006) and Liao (2013), as well as the penalties used in non-parametric and semi-nonparametric estimation; see e.g. Chen (2007), Dalalyan et al. (2006) and Green (1996).

function. Intervals where the penalty is strictly concave can be used to identify a unique preferred parameter value. As we shall see, the asymptotic properties of the PenII estimator can be established under very mild regularity conditions on the nature of the penalty function.

Third, the influence of the penalty is allowed to vanish asymptotically at any pre-specified rate. Depending on this rate, the penalty may or may not influence the asymptotic distribution of the estimator. Different rates may reflect the extent to which the researcher wishes the pre-sample information to influence the asymptotic behavior of the PenII estimator as the sample size diverges to infinity. In practice, the choice regarding the influence of the penalty can be made based on a second criterion. In this paper we explore both in-sample and out-of-sample criteria for setting the penalty strength through the use of a validation sample. The PenII estimator is also similar to a Bayesian estimator is this respect as the selection of the prior may take the sample size into account and depend on the data.

The remainder of the article is organized as follows. Section 2 introduces the PenII estimator. Section 3 establishes its asymptotic properties. Section 4 analyses finite-sample properties by means of a Monte Carlo exercise and discusses how to choose the influence of the penalty in practice. Section 5 applies the new PenII estimator to a state-of-the-art DSGE model. Section 6 concludes.

2 The Penalized Indirect Inference Estimator

Following Gourieroux et al. (1993), we let $\widehat{\boldsymbol{\beta}}_T$ denote a q-dimensional vector of auxiliary statistics describing the sample of observed data $\mathbf{x}_1, ..., \mathbf{x}_T$. Similarly, $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ is the vector of auxiliary statistics that describes the artificial data simulated from the structural model of interest. The structural model is parameterized by the p-dimensional vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. The auxiliary statistics vector $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ is defined as an average of S vectors $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ obtained from S streams $\{\widetilde{x}_{1,s}, ..., \widetilde{x}_{T,s}\}_{s=1}^{S}$ of simulated data,

$$\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) \; = \; \frac{1}{S}\sum_{s=1}^{S} \widetilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta}).$$

The penalized indirect inference (PenII) estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ is then defined as,

$$\widehat{\boldsymbol{\theta}}_{T,S} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big)' \boldsymbol{\Omega} \Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) + \pi_T(\boldsymbol{\theta}) \ ,$$

where $\pi_T : \Theta \to [0, \infty)$ is a penalty function that depends on sample size T; and $\boldsymbol{\Omega}$ is a weighting matrix. Note that the additive nature of the penalty can be generalized. In particular, the theory considered here extends naturally to the case of a multiplicative penalty function by means of a log transformation.

Consider the case of a scalar parameter $\boldsymbol{\theta} = \theta \in [0, 1]$. A trivially simple example of a penalty function is given by,

$$\pi_T(\theta) = \frac{1}{T}\Big(\theta - \frac{1}{2}\Big)^2 .$$

This quadratic penalty function penalizes deviations of $\widehat{\boldsymbol{\theta}}_T$ from the 'central' point $1/2$ of the parameter space $[0, 1]$. This penalty function vanishes at speed T asymptotically uniformly in the parameter space $[0, 1]$ since $\sup_{\theta\in[0,1]} \pi_T(\theta) = O(T^{-1})$.

Figure 1 provides simple examples of penalty functions for parameters in $\mathbb{R}$ and $\mathbb{R}^2$. Penalty (b) is similar to that in (a) but uses its explosive behavior near $\boldsymbol{\theta} = 1$ to effectively restrict the parameter space from above to values $\boldsymbol{\theta} < 1$. While penalty (c) disapproves of estimates near zero, penalty (d) excludes the region near zero altogether. Penalty (e) shows how to center the estimates at the point $\boldsymbol{\theta}^* = (1, 1)$; and penalty (f) shows how to introduce 'soft' cross-restrictions in the parameter space that penalize deviations from the relation $\theta_1 = a\theta_2$ for some $a \in \mathbb{R}$ and $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$. Darker shades represent lower values for the penalty (closer to zero), whereas lighter shades indicate larger values for the penalty. Naturally, these considerations extend to arbitrary dimensions.

Penalty functions π_T that vanish asymptotically, are interesting when the researcher wishes to impose penalties only for small sample sizes. This can be desirable when the estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ has poor finite sample properties. Poor finite sample behavior of $\widehat{\boldsymbol{\theta}}_{T,S}$ can occur, for example, when the structural parameter is weakly identified in small samples. Penalty functions are allowed to depend on the data in the same spirit as the Empirical Bayes method allows priors to be data dependent (see, e.g. Morris, 1983, and Wasserman, 2000).

The PenII estimator is also related to the mode of the pseudo-posterior distri-

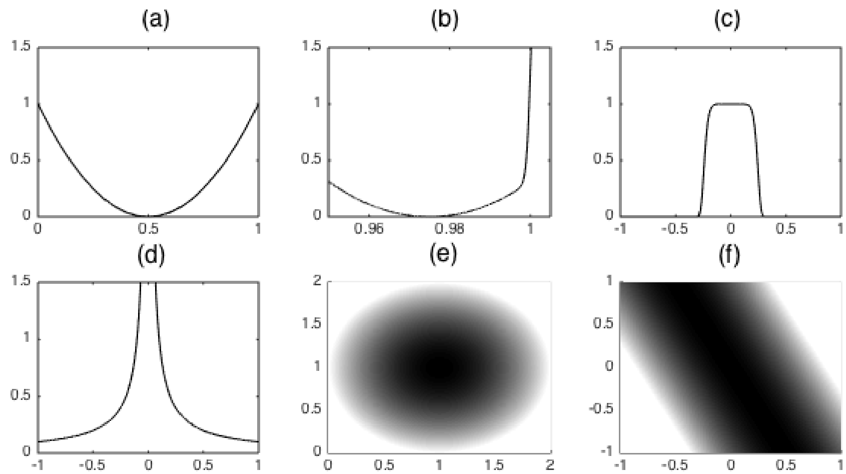

Figure 1: Examples of penalty functions: Penalty (a) and (b) center the parameter estimate at the point $\boldsymbol{\theta}^* = 0.5$. Penalties (c) and (d) penalize values close to zero. Finally, penalties (e) and (f) show how to center the estimates at a point in $\mathbb{R}^2$.

bution discussed in Chernozhukov and Hong (2003). Let $l_T(\boldsymbol{\theta}) = \prod_{t=1}^{T} f_t(\boldsymbol{\theta})$ be a pseudo-likelihood function and $\pi_T(\boldsymbol{\theta})$ be a prior. The posterior distribution $p_T(\boldsymbol{\theta})$ is then given as $p_T(\boldsymbol{\theta}) \propto l_T(\boldsymbol{\theta})\, \pi_T(\boldsymbol{\theta})$, and

$$\widehat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{T} \log l_T(\boldsymbol{\theta}) + \frac{1}{T} \log \pi_T(\boldsymbol{\theta})$$

Chernozhukov and Hong (2003) focus on priors that vanish asymptotically. As we shall see, the PenII also allows for penalties that vanish at alternative rates. Different asymptotic properties are obtained as a result of the rate at which the penalty vanishes.

3 Asymptotic Properties

In this section, we establish the asymptotic properties of the PenII estimator. When the parameter of interest is identified, then the existence, measurability, consistency and asymptotic normality of the PenII estimator can be obtained under the assumptions used in Gourieroux et al. (1993), plus a few additional conditions on the penalty function. Below, we omit the standard conditions of Gourieroux et al. (1993) and detail only the assumptions that relate to the penalty function. For convenience, the standard assumptions in Gourieroux et al. (1993) are stated in Appendix A as assumptions A.1-A.5. We also study the asymptotic properties of the PenII estimator under weak identification conditions that resemble those of Stock and Wright (2000) for the GMM estimator.

3.1 Existence and Consistency

The existence and measurability of the PenII estimator can be obtained under the standard conditions of Gourieroux et al. (1993). Let Θ denote the parameter space of the structure model. Theorem A.1 in Appendix A establishes the existence and measurability of the PenII estimator when the penalty function π_T is continuous in $\boldsymbol{\theta} \in \Theta$ for every T.

Assumption 1. *$\pi_T \in \mathbb{C}(\Theta, \mathbb{R}) \ \forall \ T \in \mathbb{N}$.*

Following Gourieroux et al. (1993), we obtain the consistency of the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ by building on the consistency of the auxiliary estimators (Assumption A.3 in Appendix A). The PenII estimator requires further that the penalty function converges asymptotically to a well defined limit function $\pi : \Theta \to [0, \infty)$.

Assumption 2. *$\sup_{\boldsymbol{\theta}\in\Theta} \|\pi_T(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ as $T \to \infty$.*

We assume that there exists a well separated unique minimizer of the limit penalized criterion function (the identifiable uniqueness of the pseudo-true parameter of interest; see e.g. White, 1980).

Assumption 3. *There exists a unique point $\boldsymbol{\theta}_0^* \in \Theta$ that minimizes the limit criterion function; i.e.*

$$\Big(\boldsymbol{b}_0 - b(\boldsymbol{\theta}_0^*)\Big)' \boldsymbol{\Omega} \Big(\boldsymbol{b}_0 - b(\boldsymbol{\theta}_0^*)\Big) + \pi(\boldsymbol{\theta}_0^*) \ < \ \Big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\Big)' \boldsymbol{\Omega} \Big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\Big) + \pi(\boldsymbol{\theta}) \quad \forall \ \boldsymbol{\theta} \in \Theta.$$

The following theorem establishes the consistency of the PenII estimator w.r.t. the pseudo-true parameter $\boldsymbol{\theta}_0^*$ that provides the best approximation of the structural model to the DGP as judged by the limit indirect inference criterion.

Theorem 1. *Let the standard indirect inference Assumptions A.1-A.3 and Assumptions 1-3 on the penalty function hold. Then $\widehat{\boldsymbol{\theta}}_{T,S} \xrightarrow{a.s.} \boldsymbol{\theta}_0^*$ as $T \to \infty \ \forall \ S \in \mathbb{N}$.*

Under an axiom of correct specification, the consistency of the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ w.r.t. the true parameter $\boldsymbol{\theta}_0$ can be obtained using the usual conditions on the injectivity of the binding function instead of the assumed uniqueness in Assumption 3; see e.g. Gourieroux et al. (1993). We formulate these conditions in Assumption A.4 in Appendix A. Of course, consistency requires also that the limit penalty function has an appropriate shape.

Assumption 4. *The penalty function* π_T *satisfies* $\pi(\boldsymbol{\theta}_0) \leq \pi(\boldsymbol{\theta})\ \forall\ \boldsymbol{\theta} \in \Theta$.

Assumption 4 describes the case where the limit penalty is well centered; i.e. the case where π has a minimum precisely at $\boldsymbol{\theta}_0 \in \Theta$. Note that the condition allows for several minima, or even a continuum of minima containing in a region of Θ containing $\boldsymbol{\theta}_0$. As we shall see, this case offers a good benchmark to analyze the various properties of the penalized estimator.

The following theorem establishes the consistency of the PenII estimator w.r.t. the true parameter $\boldsymbol{\theta}_0$.

Theorem 2. *Let the standard indirect inference assumptions A.1-A.4 and assumptions 1-4 on the penalty function hold. Then* $\widehat{\boldsymbol{\theta}}_{T,S} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $T \to \infty\ \forall\ S \in \mathbb{N}$.

3.2 Asymptotic Normality

Following Gourieroux et al. (1993), we obtain the asymptotic normality of the PenII estimator by building on the same set of conditions. These conditions are collected in Assumption A.5.

The asymptotic normality of the PenII estimator requires additionally that the penalty function converges in an appropriate manner to some limit. Assumption 5 states alternative conditions that render the PenII estimator asymptotically normal. Each condition yields a different limit result in terms of the asymptotic bias and the asymptotic variance. Throughout, we let ∇^i denote the i-th derivative operator.

Assumption 5. *The penalty function satisfies* $\pi_T \in \mathbb{C}^2(\Theta, \mathbb{R})\ \forall\ T \in \mathbb{N}$ *and one of the following conditions holds true:*

(i) $\nabla \pi_T(\boldsymbol{\theta}_0) = o(T^{-\frac{1}{2}})$ *and* $\nabla^2 \pi(\boldsymbol{\theta}_0) = o(1)$ *as* $T \to \infty$;

(ii) $\nabla \pi_T(\boldsymbol{\theta}_0) = O(T^{-\frac{1}{2}})$ *and* $\nabla^2 \pi(\boldsymbol{\theta}_0) = o(1)$ *as* $T \to \infty$;

(iii) $\nabla \pi_T(\boldsymbol{\theta}_0) = o(T^{-\frac{1}{2}})$ *and* $\nabla^2 \pi(\boldsymbol{\theta}_0) = O(1)$ *as* $T \to \infty$;

(iv) $\nabla \pi_T(\boldsymbol{\theta}_0) = O(T^{-\frac{1}{2}})$ *and* $\nabla^2 \pi(\boldsymbol{\theta}_0) = O(1)$ *as* $T \to \infty$.

The role of Assumption 5 can be easily understood by analyzing an expansion of the PenII estimator. Consider, for simplicity, the case of scalar parameters $\boldsymbol{\theta} = \theta$ and $\boldsymbol{\beta} = \beta$, then it is easy to show that

$$\sqrt{T}(\widehat{\theta}_{T,S} - \theta_0) = \frac{\nabla\widetilde{\beta}_{T,S}(\theta_0)\ \sqrt{T}\Big(\widehat{\beta}_T - \widetilde{\beta}_{T,S}(\theta_0)\Big)\ -\ \frac{1}{2}\sqrt{T}\nabla\pi_T(\theta_0)}{\nabla\widetilde{\beta}_{T,S}(\theta_0)^2\ +\ \frac{1}{2}\nabla^2\pi_T(\theta_0)\ +\ o_{a.s.}(1)}.$$

Compared to the standard II estimator, there are two additional terms in this expansion that relate to the penalty function. These are the scaled derivative of the penalty function $\frac{1}{2}\sqrt{T}\nabla\pi_T(\theta_0)$, and the term $\frac{1}{2}\nabla^2\pi_T(\theta_0)$, which is proportional to the second derivative of the penalty function. Statistical inference on $\boldsymbol{\theta}_0$ can thus be affected by both the slope and curvature of the penalty function at $\boldsymbol{\theta}_0$, even when the penalty function vanishes asymptotically. In particular, note that when the curvature vanishes asymptotically, $\nabla^2\pi_T(\theta_0) = o(1)$, and the slope vanishes sufficiently fast, $\nabla\pi_T(\theta_0) = o(T^{-\frac{1}{2}})$, then we recover the asymptotic distribution of the standard indirect inference estimator since

$$\sqrt{T}(\widehat{\theta}_{T,S} - \theta_0) = \frac{\nabla\widetilde{\beta}_{T,S}(\theta_0)\ \sqrt{T}\Big(\widehat{\beta}_T - \widetilde{\beta}_{T,S}(\theta_0)\Big)\ +\ o_{a.s.}(1)}{\nabla\widetilde{\beta}_{T,S}(\theta_0)^2\ +\ o_{a.s.}(1)}.$$

In contrast, if the curvature of the penalty function at θ_0 does not vanish asymptotically, but instead converges to some constant $\nabla^2\pi_T(\theta_0) = O(1)$ as $T \to \infty$, then the asymptotic variance of the penalized indirect inference estimator will be different from that of the standard indirect inference estimator. If the limit curvature is positive at θ_0, then the asymptotic variance of the PenII estimator will be smaller than that of the unpenalized II estimator. If the limit curvature is negative, then the asymptotic variance of the PenII estimator will be larger. Note that if the penalty's curvature $\nabla^2\pi_T(\theta_0)$ converges to a negative value that cancels out with the quadratic term $\nabla\widetilde{\beta}_{T,S}(\theta_0)^2$ as $T \to \infty$

$$\nabla\widetilde{\beta}_{T,S}(\theta_0)^2\ +\ \frac{1}{2}\nabla^2\pi_T(\theta_0) = o_{a.s.}(1),$$

then, the PenII criterion is asymptotically flat at θ_0 and, as a result, the $\sqrt{T}(\widehat{\theta}_{T,S} - \theta_0)$ diverges. Inversely, if the curvature diverges, then $\sqrt{T}(\widehat{\theta}_{T,S} - \theta_0) \xrightarrow{a.s.} 0$.

Similarly, if the penalty's slope at θ_0 does not vanish fast enough, i.e. if $\sqrt{T}\nabla\pi_T(\theta_0)$

converges to a constant, so that $\nabla\pi_T(\theta_0) = O(T^{-\frac{1}{2}})$, then an asymptotic bias is introduced and asymptotic distribution of the PenII estimator differs from that of the standard II estimator. If the scaled slope $\sqrt{T}\nabla\pi_T(\theta_0)$ diverges, then $\sqrt{T}(\widehat{\theta}_{T,S} - \theta_0)$ diverges a.s. as $T \to \infty$.

Table 1 shows the four different limiting cases for which a non-degenerate limiting distribution is well defined. These are the cases considered in Assumption 5.

Table 1: Different Limit Cases for the Asymptotic Distribution of the PenII Estimator.

	$\nabla\pi_T(\theta_0) = o(T^{-\frac{1}{2}})$	$\nabla\pi_T(\theta_0) = O(T^{-\frac{1}{2}})$
$\nabla^2\pi_T(\theta_0) = o(1)$	Standard II Asymptotics	Asymptotic Bias
$\nabla^2\pi_T(\theta_0) = O(1)$	Different Asymptotic Variance	Asymptotic Bias + Different Asymptotic Variance

Theorem 3 establishes the asymptotic normality of the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$, and describes its asymptotic mean and variance under the four different limit cases described by Assumption 5. Below we denote the asymptotic variance of $\widehat{\boldsymbol{\beta}}_T$ by $\boldsymbol{\Sigma}$ and note that the asymptotic variance of $\widetilde{\boldsymbol{\beta}}_{T,S}$ is given by $S^{-1}\boldsymbol{\Sigma}$ for any $S \in \mathbb{N}$. For notational simplicity, Theorem 3 adopts the notation

$$\boldsymbol{A} := \frac{\partial \mathbf{b}(\boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega} \frac{\partial \mathbf{b}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \quad , \quad \boldsymbol{B} := \frac{\partial \mathbf{b}(\boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}} \boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega} \frac{\partial \mathbf{b}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \ ,$$

$$\text{and} \quad \boldsymbol{\Pi} = \lim_{T\to\infty} \sqrt{T}\nabla\pi_T(\boldsymbol{\theta}_0) \quad \text{under Assumption 5[ii and iv].}$$

THEOREM 3. *Let the standard indirect inference assumptions A.1-A.5, B.1 and B.2 hold. Suppose further that assumptions 1-5 on the penalty function hold. Then* $\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\right) \stackrel{d}{\to} N(\boldsymbol{\mu}, \boldsymbol{W}) \ \forall \ S \in \mathbb{N}$ *as* $T \to \infty$ *for some* $p \times 1$ *vector* $\boldsymbol{\mu}$ *and* $p \times p$ *matrix* $\boldsymbol{W}$.

1. *If Assumption 5(i) holds then:*

$$\boldsymbol{\mu} = \mathbf{0} \quad \text{and} \quad \boldsymbol{W} = \left(1 + \frac{1}{S}\right)\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} \ .$$

2. *If Assumption 5(ii) holds then:*

$$\boldsymbol{\mu} = -\frac{1}{2}\boldsymbol{A}^{-1}\boldsymbol{\Pi} \quad \textit{and} \quad \boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\ .$$

3. *If Assumption 5(iii) holds then:*

$$\boldsymbol{\mu} = \mathbf{0} \quad \textit{and} \quad \boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\Big(\boldsymbol{A}+\frac{1}{2}\nabla^2\pi(\boldsymbol{\theta}_0)\Big)^{-1}\boldsymbol{B}\Big(\boldsymbol{A}+\frac{1}{2}\nabla^2\pi(\boldsymbol{\theta}_0)\Big)^{-1}\ .$$

4. *If Assumption 5(iv) holds then:*

$$\boldsymbol{\mu} = -\frac{1}{2}\Big(\boldsymbol{A}+\frac{1}{2}\nabla^2\pi(\boldsymbol{\theta}_0)\Big)^{-1}\boldsymbol{\Pi}$$
$$\textit{and} \quad \boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\Big(\boldsymbol{A}+\frac{1}{2}\nabla^2\pi(\boldsymbol{\theta}_0)\Big)^{-1}\boldsymbol{B}\Big(\boldsymbol{A}+\frac{1}{2}\nabla^2\pi(\boldsymbol{\theta}_0)\Big)^{-1}\ .$$

Understanding the limit behavior of the PenII estimator is useful since it reveals how statistical inference may be influenced by the introduction of the penalty. Furthermore, it can help provide some guidance for the construction of penalties that are asymptotically innocuous from an inferential perspective.

3.3 Weak Identification

In general, the penalty function of the PenII estimator can be useful for incorporating pre-sample information, attenuating estimation biases, dealing with large estimation variances that provide imprecise parameter estimates, or obtaining sensible parameter estimates in the context of structural models with specification problems. We end this section by showing that the PenII estimator can also be useful in the presence of parameters that are weakly identified.

We follow Stock and Wright (2000) in splitting the parameter vector of interest $\boldsymbol{\theta} = (\boldsymbol{\theta}^{1\prime}, \boldsymbol{\theta}^{2\prime})'$ into a sub-vector $\boldsymbol{\theta}^1$ that is well identified and a sub-vector $\boldsymbol{\theta}^2$ that is weakly identified in the sense that the II criterion is asymptotically flat in $\boldsymbol{\theta}^2$.

It is important to note that the weak identification of the structural parameter vector $\boldsymbol{\theta}$ is generally unrelated to any weak identification issues in the auxiliary model. Indeed, the auxiliary estimator $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ may well converge to a well defined limit point $\mathbf{b}(\boldsymbol{\theta})$ in $\mathcal{B}$ for every $\boldsymbol{\theta} \in \Theta$. The fact that asymptotically we may find $\mathbf{b}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_1) =$

$\mathbf{b}(\boldsymbol{\theta}^1, \boldsymbol{\theta}_2^2)$ for some $\boldsymbol{\theta}_1^2 \neq \boldsymbol{\theta}_2^2$ is a separate issue from having a set valued binding function $\mathbf{b}$ which is characteristic of identification failure in the auxiliary parameter.

Assumption 6 resembles Assumption C of Stock and Wright (2000) and makes the notion of weak identification precise in the II setting. Assumption 6 indexes the binding function by sample size and states essentially that the binding function fails to be asymptotically injective.[2]

Assumption 6. $\mathbf{b}_T(\boldsymbol{\theta}_0) - \mathbf{b}_T(\boldsymbol{\theta}) = m_1(\boldsymbol{\theta}^1) + m_{2,T}(\boldsymbol{\theta})/r_T$ *where:*

(i) m_1 *is continuous in* $\boldsymbol{\theta}^1$ *and* $m_1(\boldsymbol{\theta}^1) = 0$, $m_1(\boldsymbol{\theta}^1) \neq 0 \ \forall \ \boldsymbol{\theta}^1 \neq \boldsymbol{\theta}_0^1$;

(ii) $m_{2,T}$ *is continuous on* Θ *and* $\sup_{\boldsymbol{\theta}\in\Theta} |m_{2,T}(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta})| \to 0$ *a.s. as* $T \to \infty$;

(iii) $r_T \to \infty$ *as* $T \to \infty$.

Stock and Wright (2000) set $r_T = \sqrt{T}$ and find that under weak identification the GMM estimator of $\hat{\boldsymbol{\theta}}_{T,S}^1$ is consistent for $\boldsymbol{\theta}_0^1$, but $\hat{\boldsymbol{\theta}}_{T,S}^2$ is not consistent for $\boldsymbol{\theta}_0^2$. Lemma 1 in Appendix A shows that the same result would apply to the II estimator in the absence of a penalty function. Stock and Wright (2000) show further that both $\hat{\boldsymbol{\theta}}_{T,S}^1$ and $\hat{\boldsymbol{\theta}}_{T,S}^2$ have a non-standard asymptotic distribution as a result of the inconsistent behavior of $\hat{\boldsymbol{\theta}}_{T,S}^2$. In contrast, Theorem 3 below shows that, for any non-vanishing penalty function π_T that is well centered at $\boldsymbol{\theta}_0^2$, the PenII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ is consistent for the entire vector $\boldsymbol{\theta}_0$ and asymptotically normal with respect to the sub-vector $\boldsymbol{\theta}_0^1$ $\boldsymbol{\theta}_0$. In other words, a penalty that identifies $\boldsymbol{\theta}_0^2$, not only ensures consistency with respect to the weakly identified parameters $\boldsymbol{\theta}_0^2$, but also, the asymptotic normality with respect to well identified parameters $\boldsymbol{\theta}_0^1$. Furthermore, we find that if the penalty function links the dimensions of $\boldsymbol{\theta}_0^1$ and $\boldsymbol{\theta}_0^2$, in the sense that, $\partial^2\pi(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^1\partial\boldsymbol{\theta}^2 \neq \mathbf{0}$, then the PenII estimator is also asymptotically normal with respect to $\boldsymbol{\theta}_0^2$, despite weak identification. While the limit Gaussian distribution is preserved, the asymptotic variance of the PenII estimator under weak identification differs significantly from that obtained in Theorem 3.

For simplicity, in Theorem 4 we consider only the case of well centered penalties, i.e. penalties for which $\nabla\pi_T(\boldsymbol{\theta}_0) = \mathbf{0}$. Allowing for uncentered penalties yields essentially the same type of results as for the well identified case. In Theorem 4 we

[2] As pointed out in Stock and Wright (2000), this type of weak identification is obtained through the specification of a *drifting DGP* in the terminology of Davidson and MacKinnon (1993).

obtain asymptotic results for three different types of penalty functions. In point *(i)*, we analyze penalties that *identify only the vector* $\boldsymbol{\theta}_0^2$ since they are asymptotically flat in the dimensions of $\boldsymbol{\theta}_0^1$. In point *(ii)*, we consider penalties that *identify separately* $\boldsymbol{\theta}_0^1$ *and* $\boldsymbol{\theta}_0^2$, i.e. penalties that have zero cross-derivatives asymptotically $\nabla_{12}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$. Finally, in point *(iii)*, we consider penalties that *cross-identify* $\boldsymbol{\theta}_0^1$ *and* $\boldsymbol{\theta}_0^2$, i.e. that have non-zero cross-derivatives asymptotically. These three types of penalties lead to different asymptotic results.

THEOREM 4. *Let A.1–A.4 and 1–6 hold and suppose the penalty is well centered* $\nabla_1\pi_T(\boldsymbol{\theta}_0) = \mathbf{0}$ *and* $\nabla_2\pi_T(\boldsymbol{\theta}_0) = \mathbf{0}$ *for every* $T \in \mathbb{N}$.

(i) If $\nabla^2_{11}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$, $\nabla^2_{12}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$, $\nabla^2_{21}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$ *and* $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ *is invertible. Then* $\hat{\boldsymbol{\theta}}^2_{T,S} \xrightarrow{a.s.} \boldsymbol{\theta}_0^2$ *and* $\sqrt{T}(\hat{\boldsymbol{\theta}}^1_{T,S} - \boldsymbol{\theta}^1_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{W}^1)$ *as* $T \to \infty$ *where*

$$\boldsymbol{W}^1 = \left(1 + \frac{1}{S}\right)\boldsymbol{A}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{A}^{-1}.$$

(ii) If $\nabla^2_{12}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$, $\nabla^2_{21}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$, *and furthermore,* $\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)$ *and* $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ *are both invertible. Then* $\hat{\boldsymbol{\theta}}^2_{T,S} \xrightarrow{a.s.} \boldsymbol{\theta}_0^2$ *and* $\sqrt{T}(\hat{\boldsymbol{\theta}}^1_{T,S} - \boldsymbol{\theta}^1_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{W}^1)$ *as* $T \to \infty$ *where*

$$\boldsymbol{W}^1 = \left(1 + \frac{1}{S}\right)\left(\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)\right)^{-1}\boldsymbol{\Sigma}^*\left(\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)\right)^{-1},$$

where $\boldsymbol{B}_{ij}$ *are partitions of matrix* $\boldsymbol{B}$.

(iii) If $\nabla^2_{12}\pi(\boldsymbol{\theta}_0) \neq \mathbf{0}$, $\nabla^2_{21}\pi(\boldsymbol{\theta}_0) \neq \mathbf{0}$, *and furthermore,* $\boldsymbol{B}_{11}$ *and* $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ *are both invertible. Then* $\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{W})$ *as* $T \to \infty$ *where*

$$\boldsymbol{W} = \left(1 + \frac{1}{S}\right)\begin{bmatrix} \boldsymbol{B}_{11}\boldsymbol{\Sigma}^*\boldsymbol{B}_{11} & \boldsymbol{B}_{11}\boldsymbol{\Sigma}^*\boldsymbol{B}_{12} \\ \boldsymbol{B}_{21}\boldsymbol{\Sigma}^*\boldsymbol{B}_{11} & \boldsymbol{B}_{21}\boldsymbol{\Sigma}^*\boldsymbol{B}_{12} \end{bmatrix}.$$

Theorem 4 shows in point *(i)* that $\boldsymbol{\theta}^1_{T,S}$ and $\boldsymbol{\theta}^2_{T,S}$ are consistent as long as the penalty provides asymptotic identification to the weakly identified parameters $\boldsymbol{\theta}^2_0$; i.e. if $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ is invertible. Furthermore, point *(i)* in Theorem 4 shows also that a penalty that *identifies only the vector* $\boldsymbol{\theta}^2_0$, i.e. a penalty that satisfies $\nabla^2_{11}\pi(\boldsymbol{\theta}_0) = \mathbf{0}$, and hence is flat in the dimensions of $\boldsymbol{\theta}^1_0$, is enough to obtain the asymptotic normality of $\hat{\boldsymbol{\theta}}^1_{T,S}$. Not surprisingly, Point *(ii)* shows that if the penalty is allowed to *identify separately* $\boldsymbol{\theta}^1_0$ *and* $\boldsymbol{\theta}^2_0$, i.e. if the cross-derivatives are vanish asymptotically, then the asymptotic distribution of $\hat{\boldsymbol{\theta}}^1_{T,S}$ remains normal. Perhaps more surprising is the result in point *(iii)* that, penalties that *cross-identify* $\boldsymbol{\theta}^1_0$ *and* $\boldsymbol{\theta}^2_0$, i.e. penalties that have non-zero cross-derivatives asymptotically, render both $\boldsymbol{\theta}^1_{T,S}$ and $\boldsymbol{\theta}^2_{T,S}$ asymptotically normal. The limit joint distribution is however characterized by a singular covariance matrix.

4 Finite Sample Behavior

This section provides a Monte Carlo analysis of the finite sample behavior of the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$. In particular, it shows how the penalty function can be used to obtain several desirable properties for the estimator. Section 4.1 focuses on a small RBC model with well identified parameters. Section 4.2 studies the small sample behavior of the PenII estimator under weak identification. Section 4.3 compares the relative performance of the II and PenII estimators on a medium scale DSGE model. Finally, Section 4.4 studies alternative methods for setting the strength of the penalties in practical applications.

4.1 Small RBC Model with Well Identified Parameters

Consider a the following prototypical RBC model

$$\max \mathbb{E}_t \sum_{t=0}^{\infty} \beta^t \left(\frac{c_t^{1-\psi} - 1}{1-\psi} \right) ,$$

$$k_{t+1} = (1-\delta)k_t + \exp(z_t)k_t^{\alpha} - c_t \ ,$$

$$z_t = \rho z_{t-1} + \epsilon_t \quad , \quad \{\epsilon_t\} \sim NID(0, \sigma^2) \ ,$$

where c_t stands for consumption, k_t denotes the capital stock, z_t denotes total factor productivity, ϵ_t is normal independently distributed (NID) productivity shock, and $(\psi, \delta, \alpha, \rho, \sigma^2)$ is a vector of structural parameters. According to this model, the economic agent maximizes consumption over time subject to a capital accumulation constraint. The economy is described by an AK-production function and the level of technology follows a stationary AR(1) process with no drift. If $\psi \to 1$ and $\delta = 1$, this model has a log utility function and can be solved analytically. The following two equations characterize the solution:

$$\log k_t = \log(\alpha\beta) + \alpha \log k_{t-1} + z_{t-1}$$

$$\log c_t = \log(1 - \alpha\beta) + \alpha \log k_t + z_t$$

Below we investigate the small sample behavior of the PenII estimator by means of a Monte Carlo study. In particular, we generate $N = 1000$ streams of data of length $T = 250$ from this model under the 'true' parameter values: $\alpha = 0.33$, $\beta = 0.99$, $\psi = 1.75$, $\rho = 0.95$, $\sigma = 0.0104$. These are standard parameter values taken from Heer and Maussner (2005). We also follow the literature in fixing σ^2, ψ and focusing on the estimation of (α, β, ρ). The binding function stacks variances of the data normalized by the variance of the output and correlations of the elements of the data vector.

Figure 2 compares the finite sample distribution of the II estimator (top row) and the PenII estimator (bottom row) under correct model specification. The PenII estimator imposes a quadratic penalty that centers the estimation of α at the value 0.33, and centers the estimate of β at 0.9785. This penalty is also designed to diverge as $\beta \to 1$ so as to ensure that the estimate of β is bounded above by 1. The penalty is flat in the parameter ρ.

Since the distribution of the II estimator of α is already reasonably well centered at 0.33, the PenII estimator has a similar distribution in terms of location. However, as expected, the distribution of the PenII estimator of α appears more concentrated at the true value. Figure 2 also shows that the PenII estimator is successful in producing estimates of β that avoid the 'unacceptable' values $\beta \geq 1$. In contrast, the II estimator produces a large number of parameter estimates above 1, which are 'impossible' from an economic perspective, and render the dynamic optimization problem underlying the RBC model ill defined and invalid. The distribution of the ρ estimates are similar.

Figure 3 reveals the finite sample distribution of the II and PenII estimators under incorrect specification. This is done by generating the 'observed data' from our RBC model with a CRRA utility function with $\psi = 2$, but attempting to estimate our RBC model with log utility function obtained by setting $\psi \to 1$. Despite all other features of the model remaining intact, Figure 3 shows that even this small misspecification in the utility function leads to considerable problems for the classical II estimator. In particular, the distribution of the II estimator (in the top row),

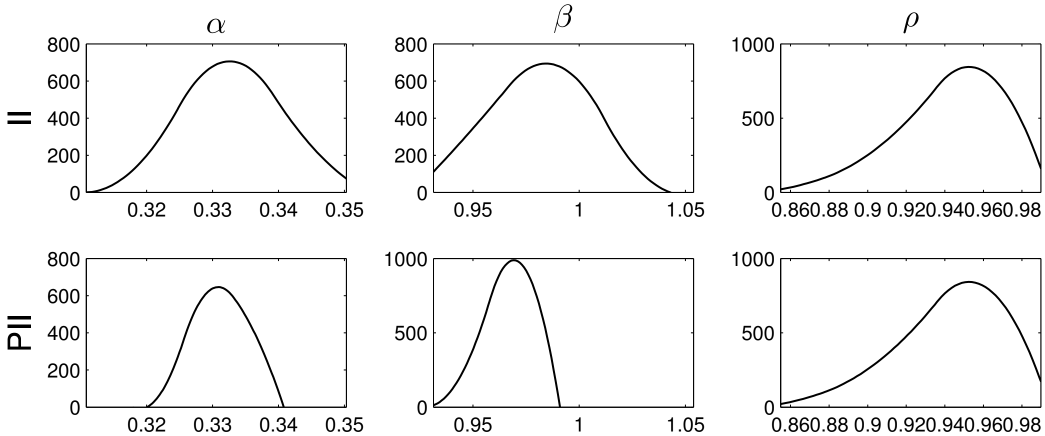

Figure 2: Distribution of II and PenII estimators for the correctly specified RBC model.

reveals that (i) the estimates of α become incomprehensible low, implying marginal productivities of capital that are hard to justify from an economic stand point; and (ii) the estimates of β become simply unacceptable from an economic theoretic stand point. It is simply impossible for an economist to justify point estimates of β that are close to zero. In comparison, the PenII estimates of both α and β are 'shifted' towards more acceptable values. Just as in a Bayesian setting, the degree to which the point estimates are shifted to acceptable values will depend on the relative strength of the penalty. Furthermore, the PenII estimate of ρ is also affected in this case although the estimate of ρ is not penalized.

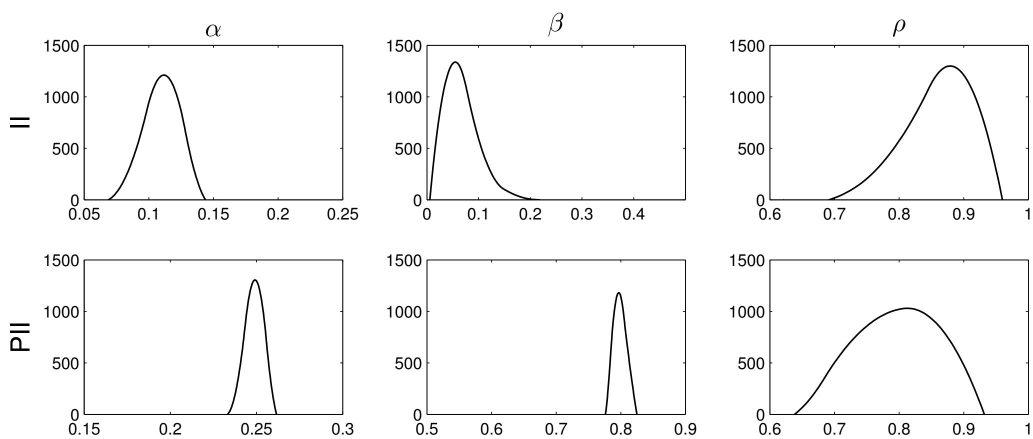

Figure 3: Distribution of II and PenII estimators for the incorrectly specified model.

In Section 3, we have noted that the effect of the penalty may vanish as the sample size increases. Theorem 3 highlighted that the penalty may also play an important role in shaping the asymptotic distribution. As such, it is important to understand the influence that the penalty may have in hypothesis testing and parameter inference when conducting statistical inference based on the asymptotic distribution of the estimator.

Figure 4 investigates the effect of the vanishing rate of the penalty function on the finite-sample and asymptotic distributions of the PenII estimator. The sample sizes are $T = 100$, 250, 500 and 1000. For simplicity, we focus on making inference about the parameter $\alpha_0 = 0.33$ while fixing the remaining parameters to their true values. In particular, we focus on testing the null hypothesis $H_0 : \alpha = 0.33$ at the 5% level.

We impose a quadratic penalty function of the form

$$\pi_T(\alpha) = c_T^2\Big(\alpha - (\alpha_0 + c_T^1)\Big)^2 \alpha_0 \ ,$$

where $\{c_T^1\}$ and $\{c_T^2\}$ are vanishing sequences. For illustrative purposes, we consider two alternative rates for the sequences. A slow-vanishing penalty where $c_T^2 = O(1)$ and $c_T^1 = O(T^{-1/2})$, and fast-vanishing penalty where $c_T^2 = O(T^{-\frac{1}{4}})$ and $\sim c_T^1 = O(T^{-1/2})$. Note that since the first and second derivatives of the penalty function are given by

$$\nabla\pi_T(\boldsymbol{\theta}_0) := \frac{\partial \pi_T(\alpha_0)}{\partial \alpha} = 2c_T^2\Big(\alpha - (\alpha_0 + c_T^1)\Big)\Big|_{\alpha_0} = 2c_T^2 c_T^1,$$

$$\text{and} \qquad \nabla^2\pi_T(\boldsymbol{\theta}_0) := \frac{\partial^2 \pi_T(\alpha_0)}{\partial \alpha^2} = 2c_T^2,$$

it follows that the fast-vanishing penalty satisfies Assumption 5(i) with $\nabla\pi_T(\boldsymbol{\theta}_0) = O(T^{-1/2})$ and $\nabla^2\pi_T(\boldsymbol{\theta}_0) = O(1)$, while the slow-vanishing penalty satisfies Assumption 9(iv) with $\nabla\pi_T(\boldsymbol{\theta}_0) = O(T^{-3/4})$ and $\nabla^2\pi_T(\boldsymbol{\theta}_0) = O(T^{-1/4})$.

As expected, the two bottom rows of Figure 4 show that the fast vanishing penalty changes the location and scale of the finite-sample distribution of PenII estimator, but leaves the asymptotic distribution unchanged. Indeed, for $T = 100$ (left column), the density of the (standardized) PenII estimator contrasts significantly with the 'corrected' distribution (in red) that is obtained using Theorem 3 to 'eliminate' the effect of the penalty. For the larger sample size of $T = 1000$, this discrepancy with respect to the corrected distribution is no longer identifiable by visual inspection of the graph. In contrast, the top two rows of Figure 4 show that the slow-vanishing penalty affects not only the small sample distribution of the PenII estimator, but

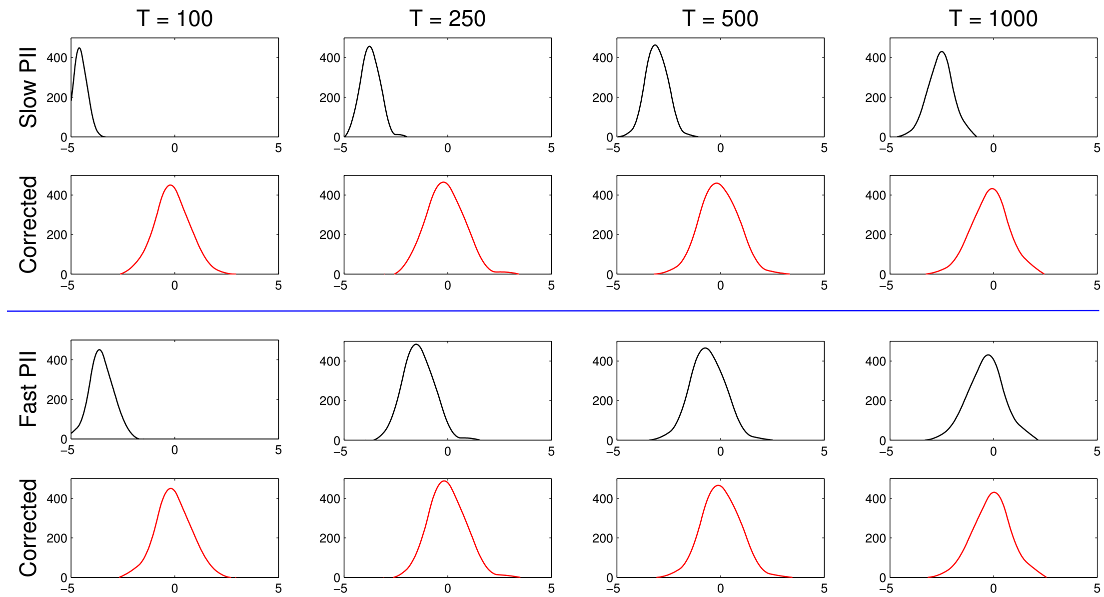

Figure 4: Corrections to the distributions. Note: This figure shows the distribution of the standardized Penalized Indirect Inference(PenII) estimator with fast and slow penalties and their corrected versions across 1000 replications. The sample sizes are 100, 250, 500 and 1000.

also, its asymptotic distribution. This is in accordance to Theorem 3, which predicts an asymptotic location and scale effect of penalties satisfying $\nabla \pi_T(\boldsymbol{\theta}_0) = O(T^{-1/2})$ and $\nabla^2 \pi_T(\boldsymbol{\theta}_0) = O(1)$.

Section S.5 in the Supplementary Appendix provides more detailed insight into the tail behavior of the distribution of the PenII estimator under the slow vanishing penalty. In particular, it reports rejection frequencies for a Wald test of nominal size of 5%, that relies on the asymptotic distributions derived in Theorem 4.

4.2 Weak Identification

We now consider the finite sample properties of II and PenII in a model that has a weakly identified parameter. We focus on the total factor productivity variable z_t and let its dynamics be described by

$$z_t = \mu_0 \frac{c}{\sqrt{T}} + \rho_0 z_{t-1} + \epsilon_t \quad , \quad \{\epsilon_t\} \sim NID(0, \sigma^2).$$

This process has a non-zero unconditional mean that *drifts* to zero, in the spirit of Stock and Wright (2000). For simplicity, we suppose that z_t is observed, and focus on the estimation of the parameter vector $\boldsymbol{\theta}_0 = (\rho_0 \; \mu_0)'$, while σ and c are assumed to be known. The vector of auxiliary statistics contains the sample mean and the sample first-order autocorrelation. The weighting matrix is set to the identity matrix.

Note that the binding function satisfies Assumption 6 and it is not asymptotically injective. In particular, it can be written as follows

$$\boldsymbol{b}_T(\boldsymbol{\theta}_0) - \boldsymbol{b}_T(\boldsymbol{\theta}) = m_1(\rho) + m_{2,T}(\theta)/r_T, \quad \text{where,}$$

$$m_1(\rho) = \rho - \rho_0, \quad m_{2,T}(\theta) = c\frac{\mu}{1-\rho} - c\frac{\mu_0}{1-\rho_0}, \quad \text{and} \;\; r_t = \sqrt{T},$$

so that ρ is identified, whereas μ is weakly identified. We use the following penalty function to illustrate how the PenII estimator may help to identify μ,

$$\pi(\boldsymbol{\theta}) = c_1(\rho - \rho_0)^2 + c_2(\mu - \mu_0)^2 + c_3(\rho - \rho_0)(\mu - \mu_0). \tag{1}$$

Note that the penalty is well centered, and the constants c_1, c_2, c_3 determine the values of the second derivatives. We set $c_1 = 0.01$, $c_2 = 0.05$, $c_3 = 0.02$. For these values of constants, all the second derivatives differ from zero at $\boldsymbol{\theta}_0$. The PenII estimator is thus expected to be asymptotically normal for both ρ and μ, by Theorem 4.

To illustrate this result, we generate $N = 10000$ samples of size $T = 250$ and $T = 1000$ and estimate $\boldsymbol{\theta}_0$ using both the II and PenII estimators. We set $\rho_0 = 0.6$, $\mu_0 = 0.5$, and $c = 0$, and $\sigma = 1$. The scaled kernel density of $\sqrt{T}(\widehat{\mu}_{T,S} - \mu_0)$ shown in Figure 5 shows that the distribution of the II estimator $\widehat{\mu}$ is not well centered around the weakly identified parameter μ_0. Furthermore, our Monte Carlo study reveals that the variance of $\sqrt{T}(\widehat{\mu}_{T,S} - \mu_0)$ increases with sample size. This suggests that the II estimator $\widehat{\mu}_{T,S}$ is at least not $\sqrt{T}$ consistent for the weakly identified parameter μ_0.

Additional Monte Carlo results are collected in Section S.1 in the Supplementary Appendix. Section S.1 reveals that the distribution of $\sqrt{T}(\widehat{\mu}_{T,S} - \mu_0)$ seems to be non-Gaussian when the cross-derivatives of the penalty function are set to zero ($c_3 = 0$). This is in accordance with Theorem 4. Furthermore, $\sqrt{T}(\widehat{\rho}_{T,S} - \rho_0)$ seems reasonably well centered around the true value of ρ_0 and the estimator seems to converge to ρ_0 as the sample size increases. This matches well with Lemma 1 in Appendix C.

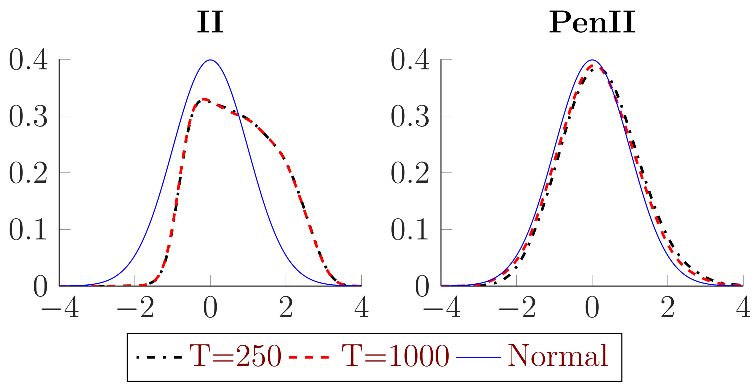

Figure 5: Scaled the kernel densities of $\sqrt{T}(\widehat{\mu}_{T,S} - \mu)$ for II and PenII estimators under weak identification of μ.

4.3 Performance of II and PenII on a Medium-Scale DSGE Model

We now study the performance of the II and PenII estimators on a medium-scale DSGE model. We consider the DSGE model introduced in Smets and Wouters (2007). For brevity, we refer to this model as SW. The SW model is a medium-scale monetary business cycle model that allows for 'sticky' prices and 'sticky' wages. The model accommodates for backward inflation indexation, habit formation in consumption, investment adjustment costs, variable capital utilization and fixed costs in production. The dynamics are driven by seven orthogonal structural shocks: total factor productivity shocks; risk premium shocks and investment-specific technology shocks, which affect the intertemporal margin; wage and price markup-shocks, which influence the intratemporal margin; and two policy shocks: exogenous spending and monetary policy shocks.

Full information maximum likelihood estimation of such models is difficult because they contain unobserved state variables, especially if nonlinear solution methods are used. Often, these models contain fewer shocks than observed state variables; in order to avoid stochastic singularities, researchers add measurement errors, but this leads to loss in estimation precision; see, amongst others, An and Schorfheide (2007), Fernandez-Villaverde (2010), and Gorodnichenko and Ng (2010) for recent discussions. Indirect inference estimation allows the researcher to avoid many of these issues. For example, Calzolari et al. (2004) propose an extension of Indirect Inference that allows the imposition of equality or inequality restrictions on the parameters of the auxiliary model. This provides a way of enhancing the information used to estimate the parameters of the simulator. Dridi et al. (2007) provide an insightful discussion of calibration within DSGE that highlights how parts of DSGE are inevitably misspecified, and introduce the Partial II estimator that is based on parts of the model deemed correctly specified. Ruge-Murcia (2012) uses simulated method of moments to estimate a DSGE model. Creel and Kristensen (2013) propose the indirect likelihood estimation that maximizes the likelihood of the auxiliary statistic.

As auxiliary statistics we use again a vector of first-order autocorrelations and the ratios of the variances to the variance of the output, as suggested, e.g., by Dave and Dejong, 2007.[3] To speed-up computations we calculate the autocovariance matrix implied by the model using the approach discussed in Gorodnichenko and Ng (2010). In particular, the moments are computed analytically making use of the VAR structure of the reduced form. We generate the data using the posterior modes reported in Smets and Wouters (2007) as the true values and use the Anderson and Moore (1985) algorithm to solve the model.

Table 2 reports Monte Carlo performance of the II and PenII estimators over $N = 500$ samples of size $T = 200$. The column labeled II reports results for the II estimator without penalties. Smets and Wouters (2007) claim that certain parameters are not identified, so their values are calibrated. The column labeled PenII_1 reports results for the PenII estimator that penalizes 9 parameters marked with asterisk that seem to be subject to weak identification according to our empirical application, with fixed $c_T = 5$ for all of them. Finally, the PenII_2 column reveals the results of the PenII estimator with quadratic loss penalties centered around the true values for all parameters.

In this setting, the PenII estimators have overall better performance. Naturally, the penalty, which is centered around the true values, reduces both the standard deviation and the bias for most of the parameters. For example, the mean of the II estimates of φ (2.752) is far from the true value (5.480). Furthermore, the II estimator has a rather large standard deviation (1.736). In contrast, both the PenII estimators are very close to the true value and have a standard deviation that is considerably smaller than that of the II estimator. Naturally, since the PenII_2 features a stronger penalty than the PenII_1 estimator, the biases and standard deviations are smaller in the PenII_2 column. Similarly to the smaller model, penalization improves the

[3] Section S.2.2 in the Supplementary Appendix shows similar results when we estimate the SW model with real data using the vector of autocovariances as proposed, e.g., by Gorodnichenko and Ng, 2010.

Table 2: Parameter Estimates of Smets and Wouters (2007) Model. Monte Carlo Evidence.

	TRUE		II	PenII_1	PenII_2		TRUE		II	PenII_1	PenII_2
γ	1.004	Mean	1.002	1.000	1.000	r_π^*	2.030	Mean	2.257	2.033	2.028
		St.dev.	0.003	0.001	0.001			St.dev.	0.366	0.033	0.012
σ_C	1.390	Mean	1.512	1.517	1.398	r_y	0.080	Mean	0.080	0.065	0.074
		St.dev.	0.253	0.238	0.015			St.dev.	0.028	0.026	0.010
σ_l^*	1.920	Mean	1.847	1.915	1.919	$r_{\Delta y}$	0.220	Mean	0.214	0.190	0.203
		St.dev.	0.695	0.022	0.007			St.dev.	0.047	0.039	0.022
λ	0.710	Mean	0.722	0.736	0.735	ρ_w	0.820	Mean	0.814	0.815	0.826
		St.dev.	0.049	0.040	0.020			St.dev.	0.064	0.046	0.019
β	0.990	Mean	0.974	0.968	0.969	μ_w	0.620	Mean	0.612	0.609	0.618
		St.dev.	0.021	0.017	0.013			St.dev.	0.100	0.074	0.018
φ^*	5.480	Mean	2.752	5.465	5.475	ρ_r	0.290	Mean	0.295	0.291	0.300
		St.dev.	1.736	0.018	0.013			St.dev.	0.083	0.069	0.020
φ_p	1.610	Mean	1.582	1.598	1.604	ρ_a	0.950	Mean	0.925	0.933	0.941
		St.dev.	0.188	0.041	0.020			St.dev.	0.071	0.033	0.015
α	0.190	Mean	0.193	0.187	0.187	ρ_b^*	0.180	Mean	0.198	0.182	0.183
		St.dev.	0.028	0.018	0.009			St.dev.	0.087	0.033	0.014
ψ^*	0.540	Mean	0.541	0.530	0.539	ρ_p	0.740	Mean	0.751	0.788	0.774
		St.dev.	0.095	0.037	0.022			St.dev.	0.133	0.111	0.038
ι_p	0.220	Mean	0.235	0.233	0.227	μ_p	0.590	Mean	0.612	0.594	0.578
		St.dev.	0.083	0.066	0.018			St.dev.	0.138	0.101	0.018
ξ_p	0.650	Mean	0.650	0.649	0.655	ρ_i	0.640	Mean	0.668	0.678	0.672
		St.dev.	0.077	0.064	0.026			St.dev.	0.047	0.021	0.013
ι_w	0.590	Mean	0.596	0.615	0.590	ρ_g^*	0.910	Mean	0.889	0.920	0.929
		St.dev.	0.190	0.143	0.014			St.dev.	0.099	0.037	0.023
ξ_w^*	0.730	Mean	0.749	0.762	0.749	ρ_{ga}	0.390	Mean	0.419	0.394	0.387
		St.dev.	0.067	0.040	0.024			St.dev.	0.095	0.061	0.017
ρ	0.810	Mean	0.850	0.869	0.859						
		St.dev.	0.046	0.034	0.028						

Monte Carlo performance of the II estimator, the PenII estimator (PenII_1) with quadratic loss penalties centered around the true values for parameters with asterisk, and the PenII estimator (PenII_2) with quadratic loss penalties centered around the true values for all parameters.

estimation precision of penalized and non penalized parameters. Also, it makes the estimates less sensitive to different starting values of the optimization routine: the stronger the penalty.

4.4 Penalty Strength Selection

Until now our analysis proceeded in the context of penalty functions whose shape was given. In practice, it may not always be obvious what the optimal shape of the penalty might be. Even if the researcher can decide on the 'center' of the penalty, it may be especially difficult to decide on an appropriate 'strength' for the penalty, as determined by the curvature of the penalty function.

In general, the penalty strength can be set by optimizing a given criterion of interest. A criterion that immediately comes to mind is simply the distance of the point estimate $\hat{\boldsymbol{\theta}}_{T,S}$ from a certain point $\boldsymbol{\theta}$ that the researcher finds 'acceptable' or 'desirable'. Criteria that depend on the data can also be considered. Obviously, this criterion cannot be the same as the II criterion since this would simply lead to the optimal penalty strength to be set to zero. Below, we consider data-dependent criteria that rely on other information about in-sample fit that is not contained in the II criterion, as well as criteria that rely on information about out-of-sample performance through the use of a validation sample. For simplicity, we use a quadratic penalty function for the PenII estimator,

$$\pi_T(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_{pre})' \operatorname{diag}(\mathbf{c}_T)\,(\boldsymbol{\theta} - \boldsymbol{\theta}_{pre})\,,$$

where $\boldsymbol{\theta}$ is the stacked vector of the parameters of interest, $\boldsymbol{\theta}_{pre}$ is the vector of the values chosen based on pre-sample information, the vector $\mathbf{c}_T \geq 0$ determines the curvature of the penalty function, and $\operatorname{diag}(\mathbf{c}_T)$ denotes a diagonal matrix with diagonal elements corresponding to $\mathbf{c}_T$

In our illustration we consider a log likelihood criterion (LL) as an example of an in-sample criterion for setting the penalty curvature $\mathbf{c}_T$, and a root mean squared forecast error ($\text{RMSE}_{\text{forecast}}$) as an out-of-sample criterion obtained through a validation sample. In practice, we expect the LL and $\text{RMSE}_{\text{forecast}}$ criteria to suggest a smaller penalty strength when the penalty is badly centered, as the introduction of the penalty will lead on average to poorer in-sample fit and poorer forecasting performance. On the other hand, we expect the LL and $\text{RMSE}_{\text{forecast}}$ criteria to suggest, a large penalty strength when the penalty is well centered, as the introduction of the penalty will lead on average to better in-sample fit and forecasting performance. At the very minimum, the use of a data-dependent criterion has the advantage of informing the researcher about potential drawbacks of adopting a given penalty function.

For illustrative purposes we focus on the estimation of two parameters of the SW model: φ and ι_w. The parameter φ denotes the steady-state elasticity of the capital adjustment cost function; the parameter ι_w stands for wage indexation, when $\iota_w = 0$ real wages do not depend on lagged inflation. We optimize the two criteria LL and $\text{RMSE}_{\text{forecast}}$ to select the elements of the vector $\mathbf{c}_T = (\mathbf{c}_{T,\varphi}\ \ \mathbf{c}_{T,\iota_w})'$.

Figure 6 plots the LL and $\text{RMSE}_{\text{forecast}}$ criteria as functions of $\mathbf{c}_{T,\varphi}$ and $\mathbf{c}_{T,\iota_w}$ for a case where the penalty function is badly centered at $\boldsymbol{\theta}_{pre} = 0.2 \times \boldsymbol{\theta}_0$. In order to highlight the expected behavior both criteria, we generate a large sample of 2000 observations. In this way we produce accurate parameter estimates and highlight the effects of a badly centered penalty. The log likelihood LL is calculated using all the 2000 observations. The $\text{RMSE}_{\text{forecast}}$ uses the first 1900 observations for the estimation of the parameters and the last 100 observations for calculating the forecast errors.

Figure 6 shows that both criteria suggest a penalty with low strength. This is expected as the penalty is badly centered and the II parameter estimates are quite accurate for $T = 2000$. Indeed, we note that LL is maximized and $\text{RMSE}_{\text{forecast}}$ is minimized for $\mathbf{c}_T$ near zero. This offers the researcher an opportunity to think carefully about the penalties that are being imposed on the parameters.

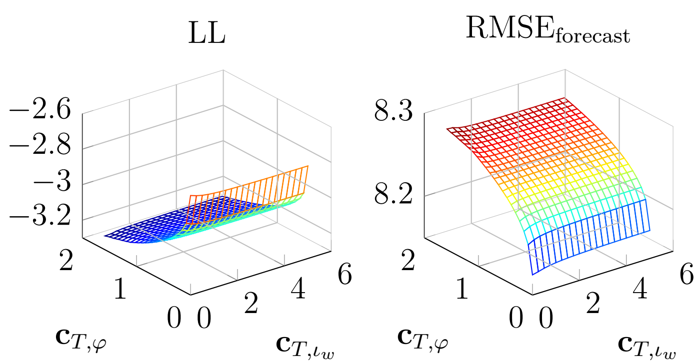

Figure 6: LL and $\text{RMSE}_{\text{forecast}}$ criteria for setting $\mathbf{c}_{T,\varphi}$ and $\mathbf{c}_{T,\iota_w}$ for a penalty that is badly centered at $\boldsymbol{\theta}_{pre} = 0.2 \times \boldsymbol{\theta}_0$.

In contrast, Figure 7 shows that when the penalty is well-centered then both the LL and the $\text{RMSE}_{\text{forecast}}$ criteria suggest large values for $\mathbf{c}_T$. This is especially true when the estimation sample is small but the validation sample is large, as the II parameter estimates will be inaccurate and benefit from a well centered penalty, and the criterion based on a large validation sample will accurately reflect the improvements produce by the well centered penalty.

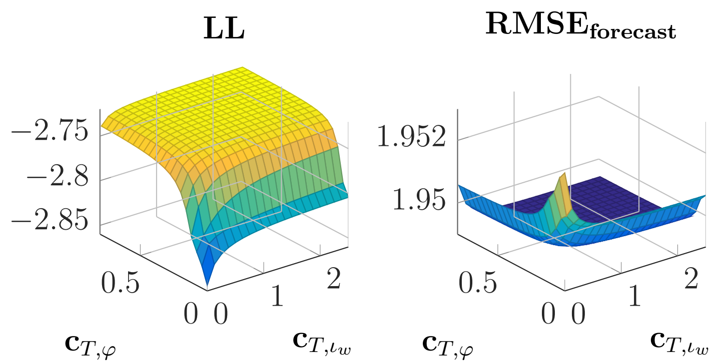

Figure 7: LL and RMSE$_{\text{forecast}}$ criteria as functions of $\mathbf{c}_{T,\varphi}$ and $\mathbf{c}_{T,\iota_w}$ for a penalty that is well centered at $\boldsymbol{\theta}_{pre} = \boldsymbol{\theta}_0$.

When both the estimation and validation samples are small, we face both large estimation uncertainty and large penalty criterion uncertainty. In small samples, the researcher may be often led to adopt an inappropriate $\mathbf{c}_T$. For example, one may be led to adopt a large $\mathbf{c}_T$ even when the penalty is badly centered. This can happen either due to the inaccuracy of the criterion LL or RMSE$_{\text{forecast}}$, or because a point estimate that overestimates $\boldsymbol{\theta}_0$ can actually benefit from a badly centered penalty that underestimates $\boldsymbol{\theta}_0$, and vice-versa.

Table 3 summarizes results for the three cases discussed above by reporting Monte Carlo frequencies of different choices of penalty strength. In Table 3, the 'Good Penalty' is correctly centered at $\boldsymbol{\theta}_0$ while the 'Bad Penalty' is centered at $0.2 \times \boldsymbol{\theta}_0$. In large samples ($T = 2000$), when the penalty is badly centered, both the LL and RMSE$_{\text{forecast}}$ criteria tend to select relatively small values for $\mathbf{c}_T$. The smaller sample size ($T = 200$) introduces more uncertainty as the estimator and the penalty criteria have less precision.same spirit, the RMSE$_{\text{forecast}}$ criterion is obtained for a small validation sample of 30 observations.

5 Empirical Application

In this section, we use the data set of Smets and Wouters (2007) to estimate the parameters of the model using II and PenII estimators.

The source of the original data is the U.S. Department of Commerce, Bureau of Economic Analysis. We take the data from the supplementary materials of Smets and Wouters (2007). The vector of observed time series consists of output, consumption, investment, real wages, hours worked, inflation, and the interest rate. Following

Table 3: Penalty Weights Suggested by Cross Validation Methods. Monte Carlo Evidence.

	'Bad Penalty' large T		'Good Penalty' large T		'Good Penalty' small T	
	φ	ι_w	φ	ι_w	φ	ι_w
			LL Criterion			
$c_T = 0$	97%	77%	1%	2%	12%	6%
$0 < c_{T,} \leq 2$	3%	18%	2%	4%	22%	7%
$2 \geq c_{T,} \leq 18$	0%	5%	16%	6%	24%	23%
$c_{T,} > 18$	0%	0%	81%	88 %	42%	64%
			$\textbf{RMSE}_{\text{forecast}}$ Criterion			
$c_T = 0$	99%	89%	9%	3%	33%	15%
$0 < c_T \leq 2$	1%	11%	14%	0%	19%	5%
$2 \geq c_T \leq 18$	0%	0%	20%	14%	7%	34%
$c_T > 18$	0%	0%	57%	83%	41%	46%

Smets and Wouters (2007), while the first four series are differenced, the last three are kept in levels. The sample size is $T = 230$.

5.1 Estimation Results

We estimate all $p = 34$ parameters of the SW model. To take into account the different scales of the variables, we adjust the curvature of the penalties for every parameter, leaving the functional form of the penalty function fixed,

$$\pi_T(\boldsymbol{\theta}) = \boldsymbol{w}_T'\boldsymbol{\kappa}(\boldsymbol{\theta}),$$

where $\boldsymbol{w}$ is a $p \times 1$ vector of weights; $\boldsymbol{\kappa}(\boldsymbol{\theta})$ is a $p \times 1$ vector, such that $\boldsymbol{\kappa}_i(\boldsymbol{\theta}) = (\theta_i - \theta_{i,SW})^2$. The weight of φ is normalized to 1, and the remainder weights are set relative to that of φ according to the relative magnitude of each parameter. Specifically, we let the elements of $\boldsymbol{w}$ be defined as the ratios of the squares of the posterior modes of Smets and Wouters (2007) to the square of φ,

$$\boldsymbol{w}_{T,i} = \frac{\varphi_{SW}^2}{\theta_{i,SW}^2} c_{T,i},$$

where $c_{T,i}$ determines the curvature of the penalty function with respect to θ_i. The sample log likelihood criteria for setting the penalty curvature $\mathbf{c}_T$ is obtained over the entire sample of 230 observations. For the out-of-sample RMSE criteria we use 200 observations to estimate the parameters and the remaining 30 observations to calculate for evaluating the RMSE.

?? reports the obtained parameter estimates. Overall, the parameter estimates obtained by the II and PenII estimators are considerably different, and they are also quite far from the posterior modes obtained by Smets and Wouters (2007). The II estimator provides the smallest estimate of α. This parameter determines the share of capital in output. Hence, the estimated value of approximately zero is hard to accept from a economic point of view. In contrast, the Smets and Wouters (2007) and PenII estimators deliver similar estimates of 0.19 and 0.15, respectively. The II estimator is also alone in suggesting almost no serial correlation in the government spending shock (values of $\rho_g \approx 0$ and $\rho_{ga} \approx 0$). In contrast, both the Smets and Wouters (2007) and PenII estimators report positive and persistent values. Interestingly, despite delivering the smallest estimate of φ, the II point estimate of ψ is very close to 1, implying that it is extremely costly to change the utilization of capital. As a result, despite the more responsive supply side of the capital market, utilization of capital is less responsive to shocks for the model implied by II estimates. In contrast, the II and PenII estimates of ρ_a are quite similar, suggesting that total factor productivity might have a unit root. The posterior mode of Smets and Wouters (2007) is 0.95 and implies considerably different dynamic behavior for productivity. II estimates of σ_a, σ_g, σ_r, σ_p, σ_w are all smaller then the values reported by Smets and Wouters (2007) and PenII.

In general, estimates of the parameters differ considerably between the II and PenII. As expected, the penalty function keeps the PenII estimates close to the posterior mode values reported by Smets and Wouters (2007). It is possible that the 'loose' II estimates of the non-identified parameters influence the estimation of the remaining parameters. Cogley (2001) reports similar findings estimating a misspecified DSGE model: a unit-root autoregressive TFP process is estimated with the stationarity restriction on the persistence parameter, as a consequence, the GMM estimate of the variance of the error term increases above the true value.

We have also conducted a detailed sensitivity analysis that describes how the PenII parameter estimates change as a function of the penalty's strength. The results can be found in Section S.2.2 of the Supplementary Appendix.

Table 4 reports the moments implied by the SW model under the parameters estimated by the different methods. It also displays how well resulting estimates fit the observed data. In particular, PenII obj fun is the value of the PenII criterion function, II obj fun is the value of the PenII criterion without penalty, and LL denotes the log-likelihood over the full sample.

Table 4: Moments Match and Objective Functions for Different Estimators

	Data	SW	II	PenII
$\text{var}(\Delta c_t)/\text{var}(\Delta y_t)$	1.42	0.61	0.98	1.22
$\text{var}(\Delta i_t)/\text{var}(\Delta y_t)$	9.32	3.63	9.25	9.32
$\text{var}(\Delta w_t)/\text{var}(\Delta y_t)$	0.44	1.08	0.74	0.43
$\text{var}(l_t)/\text{var}(\Delta y_t)$	9.54	5.05	9.55	9.54
$\text{var}(\Pi_t)/\text{var}(\Delta y_t)$	0.62	0.88	0.70	0.68
$\text{var}(r_t)/\text{var}(\Delta y_t)$	0.75	0.21	0.54	0.55
$\text{corr}(l_t, r_t)$	−0.32	−0.05	−0.27	−0.33
$\text{corr}(\Pi_t, l_{t-1})$	−0.34	−0.07	−0.33	−0.38
LL	–	−58.07	−60.26	−55.58
PenII obj fun	–	59.90	44.60	3.74
II obj fun	–	59.90	0.44	2.03

Selected moments and fit measures implied by Smets and Wouters (2007), II and PenII parameter estimates. LL_{full} denotes the log ikelihood of full sample.

The results reported in Table 4 reflect well the nature of each estimator. The posterior mode estimates reported by Smets and Wouters (2007) provide a poor approximation to a number of moments. This is natural as the estimate of Smets and Wouters (2007) are likelihood-based and did not take these specific moments as a criterion for estimation. The II estimator outperforms the other estimators in terms of the II objective function. Again, this is natural as the PenII estimator takes the influence of the penalty into account. Similarly, the PenII estimator minimizes its objective function. We further note that the selected penalties of the PenII estimator offer a good balance between all the criteria. The PenII estimator achieves the highest log likelihood, avoids the inconvenient point estimates of the II estimator for the α, ρ_g and ρ_{ga} parameters and still obtains a comparably good match for the moments.

5.2 Policy Implications

Impulse response functions (IRF) provide important tools for governments and central banks to conduct macroeconomic policy analysis. In this section, we study the IRFs implied by the different estimators for the model of SW. Figures 8 and 9 highlight the difference between the impulse responses of key macroeconomic variables to technology and government spending shocks.

Figure 8 reveals that the magnitude of the responses of output, consumption, inflation and wages is significantly smaller for II than for Smets and Wouters (2007) or PenII. This can be because the share of capital in output α is close to zero as estimated by II is considerably smaller than that obtained by II or PenII. The speed at which output converges back to equilibrium is also substantially faster for the II estimator compared to the Smets and Wouters (2007) or PenII estimators. This is a reflection of a larger estimate of the autoregressive parameter in the technological progress ρ_a found in Smets and Wouters (2007). Similarly, the inflation responses implied by all approaches are qualitatively similar, but the response implied by II is noticeably different quantitatively. II also leads to a different response in employment: producers adjust to the new state in technology by hiring less; again the effect of this shock lasts shorter for the model implied by the II estimate.

Figure 9 shows that the different estimation methods also predict different IRFs for government spending shocks. In particular, it highlights that the penalty function is crucial for obtaining a sizable reaction of aggregate investment to this kind of shock. Indeed, under the II estimates, investment is essentially non responsive. This is due to the large estimated value of ψ. The II estimate of ρ_a is close to zero, so all the effect of the shock die out quickly. Interestingly, the IRF for investments predicted by the PenII and Smets and Wouters (2007) estimators are similar in magnitude but have the opposite signs. PenII estimate of ρ_g is very close to one, and implies that the government spending is very persistent, also, the PenII estimate of σ_g is two times larger than that of Smets and Wouters (2007); therefore, consumption drops and investment rises so that production is able to match a lasting larger demand by increasing the capital.

The Supplementary Appendix contains the IRFs to other shocks of the model. For example, we find that, in response to the monetary policy shock, the II estimates produce an IRF that differs from that implied by Smets and Wouters (2007) and the PenII estimator. In particular, IRF of consumption and employment are more persistent and their magnitude is larger in the short-run. Overall, the IRFs implied by the II estimator are often an outlier.

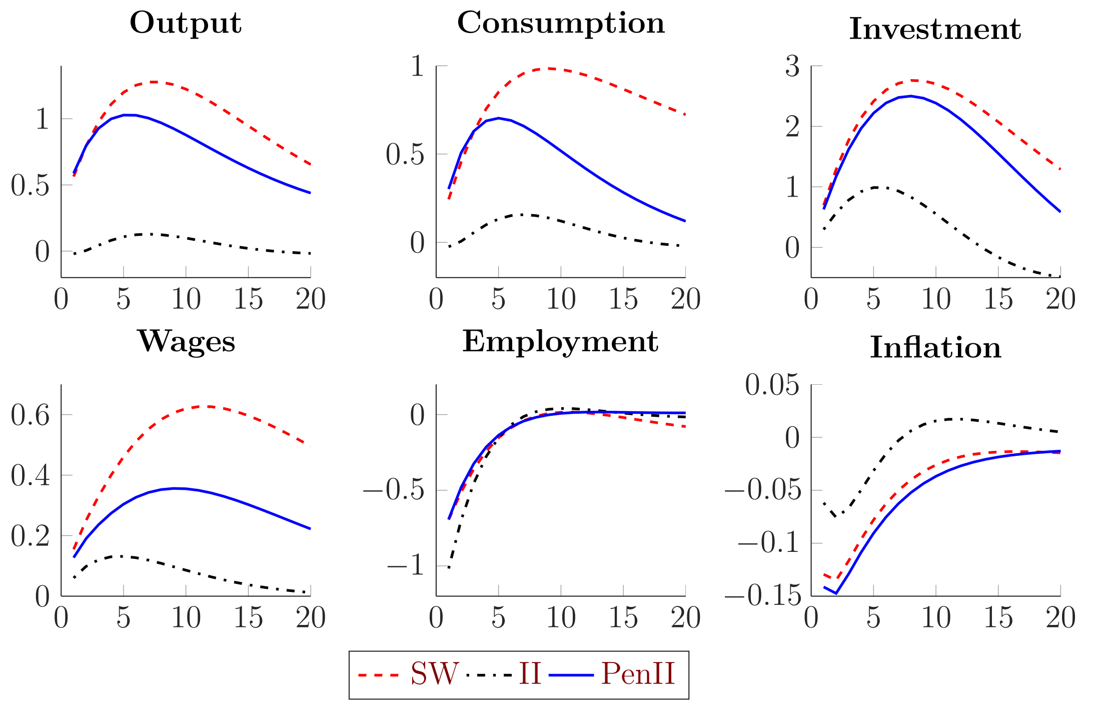

Figure 8: Estimated impulse response functions to a technology shock based on different estimators: II estimator, PenII estimator, and posterior modes from Smets and Wouters (2007) (SW).

5.3 Summary of empirical findings

Our results suggest that the PenII estimator is capable of estimating a state-of-the-art DSGE model. Both the II and PenII estimators are able to match the observed sample moments. However, the estimates produced by the II estimator are less robust than those of the PenII estimator as they are more dependent on the initial values of the parameters; see Table S.1 in the Supplementary Appendix. Furthermore, the usual II estimator fails to provide estimates that are not convincing from an economic point of view. For example, the demand for capital implied by the II estimator becomes inelastic. As a result, aggregate investment does not react to a government spending shock.

Unlike unrestricted frequentist estimation procedures such as maximum likelihood, the Bayesian posterior modes estimates reported in Smets and Wouters (2007)

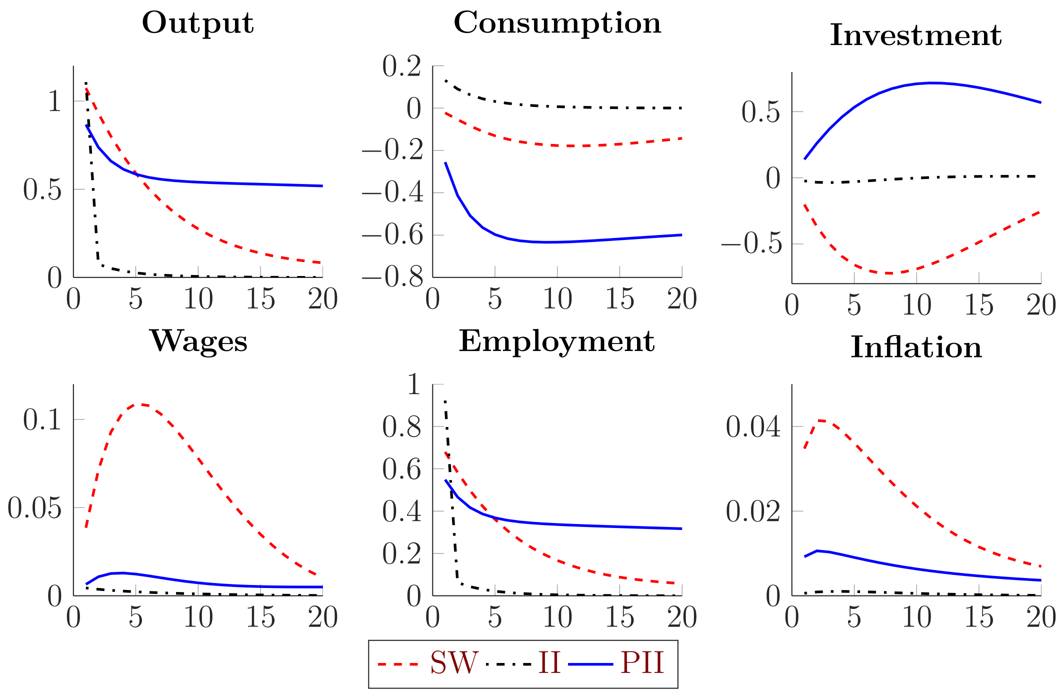

Figure 9: Estimated impulse response functions to a government spending shock based on different estimators: II estimator, PenII estimator, and posterior modes from Smets and Wouters (2007) (SW).

manage to produce estimates that are acceptable to macroeconomists. Similarly, the PenII estimator can be used to obtain parameter estimates that are acceptable at the light of economic theory and still describe reasonably well important dynamics features present in the data. Specifically, the PenII estimator is able to: (i) be considerably more robust than the II estimator in terms of criterion optimization; (ii) match well the observed sample moments; and (iii) produce parameter estimates that are compatible with economic theory and evidence.

6 Final Remarks

This paper proposed an indirect inference estimator that offers a way for the researcher to add information not included in the sample of data, and in this way, exert control over the estimator and obtain parameter estimates that are sensible from an economic perspective. The estimator was obtained by introducing a penalty in the indirect inference objective function. The penalty may vanish asymptotically or not. The asymptotic properties of the penalized indirect inference estimator were established for both correctly and incorrectly specified models, and under strong and weak parameter identification.

A Monte Carlo study revealed the role of the penalty function in shaping the finite sample distribution of the estimator. The Monte Carlo study also illustrated how the penalized indirect inference estimator may be useful in dealing with problems of parameter identification and model misspecification. Finally, the empirical application suggested that this estimator can be used for obtaining economically relevant estimates of the parameters of a state-of-the-art dynamic general equilibrium model.

The theory developed in this paper can be further extended by investigating general classes of penalties. For example, one may consider penalties that depend not only on the sample size and the structural parameters, but also, one the number of simulations or the values of the auxiliary statistics.

References

An, S. and Schorfheide, F. (2007). Bayesian Analysis of DSGE Models. *Econometric Reviews*, 26(2-4):113–172.

Anderson, G. and Moore, G. (1985). A linear algebraic procedure for solving linear perfect foresight models. *Economics letters*, 17(3):247–252.

Beyer, A. and Farmer, R. E. A. (2004). On the Indeterminacy of New-Keynesian Economics. Computing in Economics and Finance 2004 152, Society for Computational Economics.

Calzolari, G., Fiorentini, G., and Sentana, E. (2004). Constrained indirect estimation. *The Review of Economic Studies*, 71(4):945–973.

Canova, F. and Sala, L. (2009). Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, 56(4):431–449.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In Engle, R. F. and Fadden, D. M., editors, *Handbook of Econometrics*, volume 6b, chapter 76, pages 5549–5632. Elsevier.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Christiano, L. J., Eichenbaum, M., and Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, 113(1):1–45.

Cogley, T. (2001). Estimating and testing rational expectations models when the trend specification is uncertain. *Journal of Economic Dynamics and Control*, 25(10):1485–1525.

Creel, M. and Kristensen, D. (2013). Indirect likelihood inference.

Dalalyan, A. S., Golubev, G. K., and Tsybakov, A. B. (2006). Penalized maximum likelihood and semiparametric second-order efficiency. *ANN. STATIST*, pages 169–201.

Dave, C. and Dejong, D. N. (2007). *Structural Macroeconometrics*. Princeton University Press.

Davidson, R. and MacKinnon, J. G. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

Dridi, R., Guay, A., and Renault, E. (2007). Indirect inference and calibration of dynamic stochastic general equilibrium models. *Journal of Econometrics*, 136(2):397–430.

Duffie, D. and Singleton, K. J. (1990). Simulated Moments Estimation of Markov Models of Asset Prices. NBER Technical Working Papers 0087, National Bureau of Economic Research, Inc.

Dupor, B., Han, J., and Tsai, Y.-C. (2009). What do technology shocks tell us about the New Keynesian paradigm? *Journal of Monetary Economics*, 56(4):560–569.

Fernandez-Villaverde, J. (2010). The econometrics of DSGE models. *SERIEs*, 1(1):3–49.

Fukac, M. and Pagan, A. (2010). Limited information estimation and evaluation of dsge models. *Journal of Applied Econometrics*, 25(1):55–70.

Gallant, A. R. and Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, 12(04):657–681.

Gallant, A. R. and Tauchen, G. (2015). Efficient indirect bayesian estimation: With application to the dominant generalized blumenthal-getoor index of an ito semimartingale. Technical report, Working paper.

Gallant, R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

Gorodnichenko, Y. and Ng, S. (2010). Estimation of DSGE models when the data are persistent. *Journal of Monetary Economics*, 57(3):325–340.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:85–118.

Green, P. J. (1996). Penalized likelihood. In *In Encyclopedia of Statistical Sciences, Update Volume 2*, pages 578–586. John Wiley & Sons.

Heer, B. and Maussner, A. (2005). *Dynamic General Equilibrium Modelling: Computational Methods and Applications ; with 24 Tables*. Springer-Verlag GmbH.

Huber, P. (1974). *Robust Statistics*. Wiley, New York.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the berkeley symposium on mathematical statistics an.*

Lee, B.-S. and Ingram, B. F. (1991). Simulation estimation of time-series models. *Journal of Econometrics*, 47(2-3):197–205.

Liao, Z. (2013). Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory*, 29(05):857–904.

Ma, A. (2002). GMM estimation of the new Phillips curve. *Economics Letters*, 76(3):411–417.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Nason, J. M. and Smith, G. W. (2008). Identifying the new Keynesian Phillips curve. *Journal of Applied Econometrics*, 23(5):525–551.

Ruge-Murcia, F. (2007). Methods to estimate dynamic stochastic general equilibrium models. *Journal of Economic Dynamics and Control*, 31(8):2599–2636.

Ruge-Murcia, F. (2012). Estimating nonlinear DSGE models by the simulated method of moments: With an application to business cycles. *Journal of Economic Dynamics and Control*, 36(6):914 – 938.

Smets, F. and Wouters, R. (2007). Shocks and frictions in us business cycles: A Bayesian DSGE Approach. *American Economic Review*, 97(3):586–606.

Smith, A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8:563–584.

Stock, J. H. and Wright, J. H. (2000). GMM with weak identification. *Econometrica*, pages 1055–1096.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 159–180.

White, H. (1980). Nonlinear regression on cross-section data. *Econometrica: Journal of the Econometric Society*, pages 721–746.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge Books. Cambridge University Press.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

A Standard Assumptions and Theorems

Given a topological space $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$, let $\mathfrak{B}(\mathbb{A})$ denote the Borel σ-algebra generated by the topology $\mathcal{T}_{\mathbb{A}}$ of $\mathbb{A}$. We further let $(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mathbb{P}_0)$ denote the underlying complete probability space.

Assumption A.1. *$(\Theta, \mathfrak{B}(\Theta))$ and $(\mathcal{B}, \mathfrak{B}(\mathcal{B}))$ are measurable spaces and Θ is compact.*

Assumption A.2. *$\widetilde{\boldsymbol{\beta}}_{T,S} \in \mathbb{C}(\Theta, \mathcal{B})$ a.s. $\forall\ (T,S) \in \mathbb{N} \times \mathbb{N}$.*

Theorem A.1. *Let Assumptions A.1, A.2 and 1 hold. Then there exists an $\mathfrak{B}(\Theta)/\mathcal{F}$-measurable map $\widehat{\boldsymbol{\theta}}_{T,S} : \mathcal{E} \to \Theta$ satisfying*

$$\widehat{\boldsymbol{\theta}}_{T,S} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big) + \pi_T(\boldsymbol{\theta})\ \forall\ (T,S) \in \mathbb{N}^2.$$

Assumption A.3. *$\|\widehat{\boldsymbol{\beta}}_T - \boldsymbol{b}_0\| \xrightarrow{a.s.} 0$ and $\sup_{\boldsymbol{\theta}\in\Theta} \|\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) - \boldsymbol{b}(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ as $T \to \infty\ \forall\ S \in \mathbb{N}$, where $\boldsymbol{b} : \Theta \to \mathcal{B}$ is the binding function.*

Assumption A.4. *$\exists \boldsymbol{\theta}_0 \in \Theta$ such that $\boldsymbol{b}_0 = \boldsymbol{b}(\boldsymbol{\theta}_0)$ and $\boldsymbol{b} : \Theta \to \mathcal{B}$ is injective.*

Assumption A.5. *The auxiliary estimators satisfy $\widetilde{\boldsymbol{\beta}}_{T,S} \in \mathbb{C}^2(\Theta, \mathcal{B})$ a.s. $\forall\ (T,S) \in \mathbb{N} \times \mathbb{N}$ and are asymptotically normal for every $S \in \mathbb{N}$:*

(iii) $\sqrt{T}(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{b}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$ *as* $T \to \infty$;

(iv) $\sqrt{T}\big(\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - b(\boldsymbol{\theta}_0)\big) \xrightarrow{d} N(0, S^{-1}\boldsymbol{\Sigma})$ *as* $T \to \infty$.

The first two derivatives of both the auxiliary estimators and the penalty function converge uniformly to some deterministic limit for every $S \in \mathbb{N}$:

(v) $\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla^i \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) - \nabla^i \boldsymbol{b}(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ *as* $T \to \infty,\ \ i = 1, 2$;

(vi) $\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla^i \pi_T(\boldsymbol{\theta}) - \nabla^i \pi(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ *as* $T \to \infty,\ \ i = 1, 2$.

B Weak Identification Assumptions and Lemmas

Assumption B.1. *$\|\widehat{\boldsymbol{\beta}}_T - \boldsymbol{b}_T(\boldsymbol{\theta}_0)\| \xrightarrow{a.s.} 0$ and $\sup_{\boldsymbol{\theta}\in\Theta} \|\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) - \boldsymbol{b}_T(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ as $T \to \infty\ \forall\ S \in \mathbb{N}$.*

We follow Stock and Wright, 2000 in defining the estimator of the well identified parameters $\hat{\boldsymbol{\theta}}^1_{T,S}(\boldsymbol{\theta}^2)$ as a function of the weakly identified parameters

$$\widehat{\boldsymbol{\theta}}^1_{T,S}(\boldsymbol{\theta}^2) \in \arg\min_{\boldsymbol{\theta}^1} \Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)\Big)' \boldsymbol{\Omega} \Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2)\Big) + \pi_T(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2).$$

Below, the proof of consistency of $\hat{\boldsymbol{\theta}}^1_{T,S}$ is obtained by noting that $\hat{\boldsymbol{\theta}}^1_{T,S}(\boldsymbol{\theta}^2_T)$ is consistent for any sequence $\boldsymbol{\theta}^2_T$ and defining $\hat{\boldsymbol{\theta}}^1_{T,S}(\hat{\boldsymbol{\theta}}^2_{T,S})$ as in Stock and Wright, 2000.

Lemma 1. *Let assumptions A.1,A.2, B.1 and 1-3 and 6 hold. Then* $\widehat{\boldsymbol{\theta}}^1_{T,S} \xrightarrow{a.s.} \boldsymbol{\theta}^1_0$ *as* $T \to \infty \; \forall \; S \in \mathbb{N}$.

Assumption B.2. *The auxiliary estimators satisfy* $\widetilde{\boldsymbol{\beta}}_{T,S} \in \mathbb{C}^2(\Theta, \mathcal{B})$ *a.s.* $\forall \; (T,S) \in \mathbb{N} \times \mathbb{N}$ *and are asymptotically normal for every* $S \in \mathbb{N}$*:*

(iii) $\sqrt{T}(\widehat{\boldsymbol{\beta}}_T - \boldsymbol{b}_T(\boldsymbol{\theta}_0)) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$ *as* $T \to \infty$*;*

(iv) $\sqrt{T}\big(\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - b_T(\boldsymbol{\theta}_0)\big) \xrightarrow{d} N(0, S^{-1}\boldsymbol{\Sigma})$ *as* $T \to \infty$.

The first two derivatives of both the auxiliary estimators and the penalty function converge uniformly to some deterministic limit for every $S \in \mathbb{N}$*:*

(v) $\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^i \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) - \nabla^i \boldsymbol{b}_T(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ *as* $T \to \infty, \;\; i = 1,2$*;*

(vi) $\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^i \pi_T(\boldsymbol{\theta}) - \nabla^i \pi(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0$ *as* $T \to \infty, \;\; i = 1,2$.

C Proofs

Proof of Theorem 1

Following a classical consistency argument (found, e.g., in White, 1994, Theorem 3.4 or Theorem 3.3 in Gallant and White, 1988), we obtain $\widehat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}^*_0$ from the uniform convergence of the criterion function $Q_{T,S}(\boldsymbol{\theta}) := \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big) + \pi_T(\boldsymbol{\theta})$ to the limit $Q_\infty(\boldsymbol{\theta}) := \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big) + \pi(\boldsymbol{\theta})$ as $T \to \infty$,

$$\sup_{\boldsymbol{\theta} \in \Theta} |Q_{T,S}(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \; \forall \; S \in \mathbb{N} \;\; \text{as} \;\; T \to \infty \tag{2}$$

and the identifiable uniqueness of the maximizer $\boldsymbol{\theta}_0 \in \Theta$ introduced in White (1994),

$$\sup_{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon} Q_\infty(\boldsymbol{\theta}) > Q_\infty(\boldsymbol{\theta}_0) \; \forall \; \epsilon > 0. \tag{3}$$

The uniform convergence of the criterion follows from the uniform convergence of the auxiliary statistics $\widehat{\boldsymbol{\beta}}_T$ and $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ in Assumption A.3 and the uniform convergence of $\pi_T(\boldsymbol{\theta})$ in Assumption 2,

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta} |Q_{T,S}(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta} \in \Theta} &\Big| \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big) - \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big) \Big| \\ &+ \sup_{\boldsymbol{\theta} \in \Theta} \Big| \pi_T(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) \Big| = o_{a.s.}(1) + o_{a.s.}(1) = o_{a.s.}(1), \end{aligned}$$

where $\sup_{\boldsymbol{\theta} \in \Theta} \Big| \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big) - \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big) \Big| = o_{a.s.}(1)$ follows by Assumption A.3 and $\sup_{\boldsymbol{\theta} \in \Theta} \big| \pi_T(\boldsymbol{\theta}) - \pi(\boldsymbol{\theta}) \big| = o(1)$ holds by Assumption 2.

The identifiable uniqueness of $\boldsymbol{\theta}^*_0$ (see, e.g., White, 1994) follows from the assumed uniqueness of $\boldsymbol{\theta}^*_0$ in Assumption 3, the continuity of the limit criterion Q_∞ on Θ (implied by the uniform convergence of $\{Q_T\}$ and the Arzella-Ascolli Theorem) and the compactness of Θ.

Proof of Theorem 2

We follow again the classical argument found e.g. in (White, 1994, Theorem 3.4) or Theorem 3.3 in Gallant and White (1988). The uniform convergence of the criterion function

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_{T,S}(\boldsymbol{\theta}) - Q_{\infty}(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \;\forall\; S \in \mathbb{N} \;\text{ as }\; T \to \infty \tag{4}$$

is again obtained by the same argument as in the Proof of Theorem 1.

The identifiable uniqueness of the maximizer $\boldsymbol{\theta}_0 \in \Theta$ follows from the continuity of the limit criterion Q_∞ on Θ and, compactness of Θ and the uniqueness of $\boldsymbol{\theta}_0$ as a maximizer of Q_∞, i.e. $Q_\infty(\boldsymbol{\theta}_0) < Q_\infty(\boldsymbol{\theta}) \;\forall\; \boldsymbol{\theta} \in \Theta$. The compactness of Θ is directly assumed. The continuity of Q_∞ is obtained by the same argument as in the proof of Theorem 1. The uniqueness of $\boldsymbol{\theta}_0$ follows from Assumptions A.4 and 4. In particular, note that Assumption A.4 ensures that $\boldsymbol{\theta}_0$ minimizes the quadratic term $\big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)'\boldsymbol{\Omega}\big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big) = \big(b(\boldsymbol{\theta}_0) - \boldsymbol{b}(\boldsymbol{\theta})\big)'\boldsymbol{\Omega}\big(b(\boldsymbol{\theta}_0) - \boldsymbol{b}(\boldsymbol{\theta})\big)$ by setting it to zero. Assumption 4 (i) and (ii) ensure that the limit penalty $\pi : \Theta \to [0, \infty)$ has a minimum at $\boldsymbol{\theta}_0$. As a result, Q_∞ is uniquely minimized at $\boldsymbol{\theta}_0$. Finally, under Assumption 4 (iii), the penalty vanishes uniformly from limit criterion, and hence $Q_\infty(\boldsymbol{\theta}) = \big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)'\boldsymbol{\Omega}\big(\boldsymbol{b}_0 - \boldsymbol{b}(\boldsymbol{\theta})\big)$ and the uniqueness of $\boldsymbol{\theta}_0$ follows from Assumption A.4.

Proof of Theorem 3

Following Gourieroux et al. (1993), we note that the first order condition for the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ is given by

$$-2\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\widehat{\boldsymbol{\theta}}_{T,S})}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\widehat{\boldsymbol{\theta}}_{T,S})\Big) + \nabla\pi_T(\widehat{\boldsymbol{\theta}}_{T,S}) = \mathbf{0}$$

where $\mathbf{0}$ denotes a vector of zeros. Application of a mean value theorem at $\boldsymbol{\theta}_0$ yields

$$\begin{aligned}\sqrt{T}(\widehat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0) \;=\; &\left[\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}} - \frac{\partial^2\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}'\boldsymbol{\Omega}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^*_{T,S})\Big) + \frac{1}{2}\nabla^2\pi_T(\boldsymbol{\theta}^*_{T,S})\right]^{-1}\\ &\times\left(\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\Big) - \frac{1}{2}\sqrt{T}\nabla\pi_T(\boldsymbol{\theta}_0)\right).\end{aligned}$$

The consistency of $\widehat{\boldsymbol{\theta}}_T$ and the uniform convergence of $\nabla^2\widetilde{\beta}_{T,S}$ as $T \to \infty$ ensures that

$$-\frac{\partial^2\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}'\boldsymbol{\Omega}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^*_{T,S})\Big) = o_{a.s.}(1) \quad \text{as} \quad T \to \infty.$$

Furthermore, the uniform convergence of $\nabla\widetilde{\beta}_{T,S}$ and $\nabla^2\pi_T$ ensures also that

$$\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}^*_{T,S})}{\partial\boldsymbol{\theta}} - \frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = o_{a.s.}(1) \quad \text{as} \quad T \to \infty\;,$$

$$\text{and} \quad \nabla^2 \pi_T(\boldsymbol{\theta}^*_{T,S}) - \nabla^2 \pi_T(\boldsymbol{\theta}_0) = o(1) \quad \text{as} \quad T \to \infty.$$

As a result, we can re-write the expression for $\sqrt{T}(\widehat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)$ as

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0) = \left[\frac{\partial \widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}' \boldsymbol{\Omega} \frac{\partial \widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{1}{2} \nabla^2 \pi_T(\boldsymbol{\theta}_0) + o_{a.s.} \right]^{-1}$$
$$\times \left(\frac{\partial \widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}' \boldsymbol{\Omega} \sqrt{T} \Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) \Big) - \frac{1}{2} \sqrt{T} \nabla \pi_T(\boldsymbol{\theta}_0) \right).$$

The expressions for $\boldsymbol{\mu}$ and $\boldsymbol{W}$ now follow immediately from the fact that Assumption A.5(iii) and (iv) imply for every $S \in \mathbb{N}$, and as $T \to \infty$, the weak convergence

$$\sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\Big) = \sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - b(\boldsymbol{\theta}_0)\Big) - \sqrt{T}\Big(\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - b(\boldsymbol{\theta}_0)\Big) \;\; \xrightarrow{d} \;\; N\Big(\mathbf{0}, (1 + S^{-1})\boldsymbol{\Sigma}\Big).$$

Proof of Lemma 1

For any given sequence $\{\boldsymbol{\theta}^2_T\}$ in Θ, we obtain $\widehat{\boldsymbol{\theta}}^1_T(\boldsymbol{\theta}^2_T) \xrightarrow{a.s.} \boldsymbol{\theta}^1_0$ from the uniform convergence of the criterion function $Q_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_T) := \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_T)\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_T)\big)$ to the limit $Q_\infty(\boldsymbol{\theta}^1) := m_1(\boldsymbol{\theta}^1)' \boldsymbol{\Omega} m_1(\boldsymbol{\theta}^1)$ as $T \to \infty$, and the identifiable uniqueness of the maximizer $\boldsymbol{\theta}^1_0 \in \Theta$ introduced in White (1994),

$$\sup_{\boldsymbol{\theta}^1 \in \Theta^1 : \|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^1_0\| > \epsilon} Q_\infty(\boldsymbol{\theta}^1) > Q_\infty(\boldsymbol{\theta}^1_0) \; \forall \, \epsilon > 0. \tag{5}$$

The uniform convergence over Θ^1 follows from the uniform convergence of the auxiliary statistics $\widehat{\boldsymbol{\beta}}_T$ and $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2_T)$ over Θ^1 for any sequence $\{\boldsymbol{\theta}^2_T\}$; see Section S.4.2 in the Supplementary Appendix for details. The identifiable uniqueness of $\boldsymbol{\theta}^1_0$ follows from the uniqueness of $\boldsymbol{\theta}^1_0$ and the continuity of the limit criterion Q_∞ in $\boldsymbol{\theta}^1$ under Assumption 6.

Proof of Theorem 4

We obtain $\widehat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ from the uniform convergence of the criterion function $Q_{T,S}(\boldsymbol{\theta}) := \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big)' \boldsymbol{\Omega} \big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\big) + \pi_T(\boldsymbol{\theta})$ to the limit $Q_\infty(\boldsymbol{\theta}) := m_1(\boldsymbol{\theta}^1)' \boldsymbol{\Omega} m_1(\boldsymbol{\theta}^1) + \pi(\boldsymbol{\theta})$ as $T \to \infty$, and the identifiable uniqueness of the maximizer $\boldsymbol{\theta}^1_0 \in \Theta$ introduced in White (1994),

$$\sup_{\boldsymbol{\theta}^1 \in \Theta^1 : \|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^1_0\| > \epsilon} Q_\infty(\boldsymbol{\theta}^1) > Q_\infty(\boldsymbol{\theta}^1_0) \; \forall \, \epsilon > 0. \tag{6}$$

The uniform convergence of the criterion over Θ follows from the uniform convergence of the auxiliary statistics $\widehat{\boldsymbol{\beta}}_T$ and $\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ over Θ; see section S.4.3 in the Supplementary Appendix for details. The identifiable uniqueness of $\boldsymbol{\theta}_0$ follows from the uniqueness of $\boldsymbol{\theta}_0$ and the continuity of the limit criterion Q_∞ in $\boldsymbol{\theta}$ under Assumption 6.

Let p_1 and p_2 denote the number of elements in $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ respectively. By construction $p_1 + p_2 = p$. The asymptotic normality of $\hat{\boldsymbol{\theta}}^1_T$ follows by noting that the first order condition for the PenII estimator $\widehat{\boldsymbol{\theta}}_{T,S}$ is given by

$$-2\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\widehat{\boldsymbol{\theta}}_{T,S})}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\Big(\widehat{\boldsymbol{\beta}}_T-\widetilde{\boldsymbol{\beta}}_{T,S}(\widehat{\boldsymbol{\theta}}_{T,S})\Big)+\nabla\pi_T(\widehat{\boldsymbol{\theta}}_{T,S})=\mathbf{0}$$

where $\mathbf{0}$ denotes a vector of zeros. Following the proof of Theorem 3 we arrive at

$$\begin{aligned}\sqrt{T}(\widehat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0) \;&=\left[\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}+\frac{1}{2}\nabla^2\pi_T(\boldsymbol{\theta}_0)\right]^{-1}\\ &\qquad\times\left(\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T-\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\Big)-\frac{1}{2}\sqrt{T}\nabla\pi_T(\boldsymbol{\theta}_0)\right).\end{aligned}$$

Now, given Assumptions B.1 and 6, we have that

$$\left\|\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'-\frac{\partial b_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\right\|\overset{a.s.}{\to}0\quad\text{and}\quad\frac{\partial b_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\to\begin{bmatrix}\frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1}'\\ \mathbf{0}_{(p_2\times q)}\end{bmatrix}\quad\text{as } T\to\infty.$$

This implies that

$$\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\quad\overset{a.s.}{\to}\quad\begin{bmatrix}\frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1}'\\ \mathbf{0}_{(p_2\times q)}\end{bmatrix}\quad\text{as } T\to\infty,$$

where $\mathbf{0}_{(p_2\times q)}$ denotes a $p_2\times q$ matrix of zeros. Hence, we have by application of the continuous mapping theorem,

$$\left[\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}'\boldsymbol{\Omega}\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}+\frac{1}{2}\nabla^2\pi_T(\boldsymbol{\theta}_0)\right]^{-1}\quad\overset{a.s.}{\to}\quad\begin{bmatrix}\boldsymbol{A}+\frac{1}{2}\nabla_{11}^2\pi(\boldsymbol{\theta}_0) & \frac{1}{2}\nabla_{12}^2\pi(\boldsymbol{\theta}_0)\\ \frac{1}{2}\nabla_{21}^2\pi(\boldsymbol{\theta}_0) & \frac{1}{2}\nabla_{22}^2\pi(\boldsymbol{\theta}_0)\end{bmatrix}^{-1}$$

as $T\to\infty$, where $\boldsymbol{A}$ is a $p_1\times p_1$ matrix and $\nabla_{ij}^2\pi(\boldsymbol{\theta}_0)$ is a $p_i\times p_j$ matrix, for $i=1,2$ and $j=1,2$, defined as

$$\boldsymbol{A}:=\frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1}'\boldsymbol{\Omega}\frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1}\quad\text{and}\quad\nabla_{ij}^2\pi(\boldsymbol{\theta}_0):=\frac{\partial^2\pi(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^i\partial\boldsymbol{\theta}^j}.$$

We note further that the limit inverse matrix is a $p\times p$ matrix given by

$$\begin{bmatrix}\boldsymbol{A}+\frac{1}{2}\nabla_{11}^2\pi(\boldsymbol{\theta}_0) & \frac{1}{2}\nabla_{12}^2\pi(\boldsymbol{\theta}_0)\\ \frac{1}{2}\nabla_{21}^2\pi(\boldsymbol{\theta}_0) & \frac{1}{2}\nabla_{22}^2\pi(\boldsymbol{\theta}_0)\end{bmatrix}^{-1}=\begin{bmatrix}\boldsymbol{B}_{11} & \boldsymbol{B}_{12}\\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22}\end{bmatrix}\tag{7}$$

where

$$\boldsymbol{B}_{11}:=\left(\boldsymbol{A}+\frac{1}{2}\nabla_{11}^2\pi(\boldsymbol{\theta}_0)-\frac{1}{2}\nabla_{12}^2\pi(\boldsymbol{\theta}_0)\big(\nabla_{22}^2\pi(\boldsymbol{\theta}_0)\big)^{-1}\nabla_{21}^2\pi(\boldsymbol{\theta}_0)\right)^{-1},$$

$$\boldsymbol{B}_{12}:=-\boldsymbol{B}_{11}\nabla_{12}^2\pi(\boldsymbol{\theta}_0)\big(\nabla_{22}^2\pi(\boldsymbol{\theta}_0)\big)^{-1},\quad\boldsymbol{B}_{21}:=-\big(\nabla_{22}^2\pi(\boldsymbol{\theta}_0)\big)^{-1}\nabla_{21}^2\pi(\boldsymbol{\theta}_0)\boldsymbol{B}_{11},$$

$$\boldsymbol{B}_{22} := \big(\frac{1}{2}\nabla^2_{22}\pi(\boldsymbol{\theta}_0)\big)^{-1} - \big(\nabla^2_{22}\pi(\boldsymbol{\theta}_0)\big)^{-1}\nabla^2_{21}\pi(\boldsymbol{\theta}_0)\boldsymbol{B}_{12}\ .$$

Finally, by Assumption B.2, since

$$\begin{aligned}
&\sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\Big)\\
&\quad = \sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - b_T(\boldsymbol{\theta}_0)\Big) - \sqrt{T}\Big(\widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - b_T(\boldsymbol{\theta}_0)\Big) \ \ \overset{d}{\to}\ \ N\Big(\mathbf{0}_{(p\times 1)}, (1+S^{-1})\boldsymbol{\Sigma}\Big).
\end{aligned}$$

we obtain

$$\frac{\partial\widetilde{\boldsymbol{\beta}}_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}' \boldsymbol{\Omega}\sqrt{T}\Big(\widehat{\boldsymbol{\beta}}_T - \widetilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\Big) \ \ \overset{a.s.}{\to}\ \ N\Big(\mathbf{0}_{(p\times 1)},\ \Sigma^*\Big)$$

$$\text{where}\quad \boldsymbol{\Sigma}^* := \begin{bmatrix} \frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1}' \\ \mathbf{0}_{(p_2\times q)} \end{bmatrix} \boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{\Omega} \begin{bmatrix} \frac{\partial m_1(\boldsymbol{\theta}_0^1)}{\partial\boldsymbol{\theta}^1} & \mathbf{0}_{(q\times p_2)} \end{bmatrix}.$$

As a result, we conclude that, in general, $\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0) \overset{d}{\to} N(\boldsymbol{\mu}, \mathbf{W})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{B}_{11}\boldsymbol{\Pi}_1 + \boldsymbol{B}_{12}\boldsymbol{\Pi}_2 \\ \boldsymbol{B}_{21}\boldsymbol{\Pi}_1 + \boldsymbol{B}_{22}\boldsymbol{\Pi}_2 \end{bmatrix} \quad \text{where}\quad \boldsymbol{\Pi}_i := -\lim_{T\to\infty}\sqrt{T}\frac{\partial\pi_T(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}^i,\ i=1,2\ ,$$

$$\text{and}\quad \boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\begin{bmatrix} \boldsymbol{B}_{11}\boldsymbol{\Sigma}^*\boldsymbol{B}_{11} & \boldsymbol{B}_{11}\boldsymbol{\Sigma}^*\boldsymbol{B}_{12} \\ \boldsymbol{B}_{21}\boldsymbol{\Sigma}^*\boldsymbol{B}_{11} & \boldsymbol{B}_{21}\boldsymbol{\Sigma}^*\boldsymbol{B}_{12} \end{bmatrix}.$$

This yields the desired result. Namely, if $\nabla^2_{11}\pi(\boldsymbol{\theta}_0) = \mathbf{0}_{(p_1\times p_1)}$, $\nabla^2_{12}\pi(\boldsymbol{\theta}_0) = \mathbf{0}_{(p_1\times p_2)}$, $\nabla^2_{21}\pi(\boldsymbol{\theta}_0) = \mathbf{0}_{(p_2\times p_1)}$, and furthermore, $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ is invertible, then the inverse matrix in eq. (7) is well defined and

$$\boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{A}^{-1} & \mathbf{0}_{(p_1\times p_2)} \\ \mathbf{0}_{(p_2\times p_1)} & \Big(\frac{1}{2}\nabla^2_{22}\pi(\boldsymbol{\theta}_0)\Big)^{-1} \end{bmatrix}.$$

Alternatively, if $\nabla^2_{12}\pi(\boldsymbol{\theta}_0) = \mathbf{0}_{(p_1\times p_2)}$, $\nabla^2_{21}\pi(\boldsymbol{\theta}_0) = \mathbf{0}_{(p_2\times p_1)}$, and furthermore, $\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)$ and $\nabla^2_{22}\pi(\boldsymbol{\theta}_0)$ are both invertible, then the inverse matrix in eq. (7) is well defined and

$$\boldsymbol{W} = \Big(1+\frac{1}{S}\Big)\begin{bmatrix} \Big(\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)\Big)^{-1}\boldsymbol{\Sigma}^*\Big(\boldsymbol{A} + \frac{1}{2}\nabla^2_{11}\pi(\boldsymbol{\theta}_0)\Big)^{-1} & \mathbf{0}_{(p_1\times p_2)} \\ \mathbf{0}_{(p_2\times p_1)} & \Big(\frac{1}{2}\nabla^2_{22}\pi(\boldsymbol{\theta}_0)\Big)^{-1} \end{bmatrix}.$$