Parity Regression Estimation*

Vali Asimit[1] Ziwei Chen[1] Bogdan Ichim[2,3] Pietro Millossovich[1]

Abstract

Multiple linear regression is one of the most widely used predictive models across a broad range of applications. We propose a novel regression framework that, instead of minimising the aggregate prediction error in the dependent variable, distributes the total prediction error evenly across all parameters. This approach is particularly suitable for data affected by substantial noise, as is often the case for time series data where structural changes and evolving trends are common. We provide a theoretical characterisation of our proposed estimator, named *Parity Regression*. Its properties are compared with those of existing penalised and shrinkage estimators in the literature. Both synthetic experiments and real-data applications demonstrate that the theoretical guarantees of the proposed method are reflected in practice.

Keywords: Ordinary Least Square, Parity, Ridge Regression, Shrinkage Estimation
JEL classification:

*Ziwei Chen is corresponding author. (1) Bayes Business School, City St George's, University of London, 106 Bunhill Row, London, EC1Y 8TZ, UK. (2) University of Bucharest, Faculty of Mathematics and Computer Science, Str. Academiei 14, 010014 Bucharest, Romania. (3) Simion Stoilow Institute of Mathematics of the Romanian Academy, Research Unit 5, C.P. 1-764, 010702 Bucharest, Romania. Email addresses: asimit@city.ac.uk, ziwei.chen.3@citystgeorges.ac.uk, Pietro.Millossovich.1@citystgeorges.ac.uk, bogdan.ichim@fmi.unibuc.ro, bogdan.ichim@imar.ro.
R Package implementing our estimators are available at https://github.com/Ziwei-ChenChen/savvyPR.

1 Introduction

Multiple linear regression is one of the most widely used predictive and inferential models across a broad range of scientific disciplines, including economics, engineering, medicine, and the social sciences. The model relates a scalar response random variable Y to a set of explanatory random variables $X_1, X_2, \ldots, X_p$ through the linear model

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p + \varepsilon,$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)^\top$ is the unknown parameter vector and ε is a random error term with $\mathrm{E}[\varepsilon] = 0$. For a sample of size n drawn from $(Y, \mathbf{X})$, let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ denote the observed vector of responses and $\widetilde{\mathbf{X}}$ be the $n \times (p+1)$ design matrix with rows $\tilde{\mathbf{x}}_i^\top$, where $\tilde{\mathbf{x}}_i = (1, \mathbf{x}_i^\top)^\top$ for all $1 \leq i \leq n$.

A common objective in regression analysis is to estimate $\boldsymbol{\theta}$ via estimators with a low *mean squared error (MSE)*; the MSE of an estimator is explicitly defined in (3). This is typically achieved by minimising a loss functional $\mathcal{L}$ measuring the discrepancy between the dependent variable and its linear predictor. The standard choice is the l_2 loss, which leads to the well-known *Ordinary Least Squares (OLS)* estimator, denoted as $\widehat{\boldsymbol{\theta}}^{\mathrm{OLS}}$, which is obtained by minimising the *residual sum of squares (RSS)*:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{OLS}} := \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \, \mathcal{RSS}(\boldsymbol{\theta}), \quad \text{where} \quad \mathcal{RSS}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\theta}^\top \tilde{\mathbf{x}}_i - y_i)^2. \tag{1}$$

Here, $\mathcal{L} = \mathcal{RSS}$. The OLS estimator has a closed-form solution as follows:

$$\widehat{\boldsymbol{\theta}}^{\mathrm{OLS}} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \mathbf{y}, \tag{2}$$

provided that $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$ is invertible. Note that $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$ is a symmetric matrix that is always positive semi-definite and not necessarily positive definite, which would guarantee the existence of $(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})^{-1}$. The OLS estimator is often called the *Best Linear Unbiased Estimator (BLUE)* according to the *Gauss-Markov Theorem* (Gauss, 1821; Markov, 1912), which guarantees it

has the lowest variance among all unbiased linear estimators. Furthermore, if the error term is normally distributed, OLS coincides with the maximum likelihood estimator, allowing exact finite-sample inference (Seber and Lee, 2003).

A limitation of (2) is that its estimation error is driven by the estimation error of $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$, which can be highly problematic since the empirical eigenvalues of a matrix are often poor estimators of the population eigenvalues; for a detailed discussion, see Asimit et al. (2026b), with a summary provided in Section 1.1. When the sample is affected by substantial noise – as is common in time series data with structural changes or evolving trends – the *out-of-sample (OOS)* performance of OLS deteriorates further. This motivates the need for a more robust linear regression estimator suitable for such settings.

In this paper, we introduce our novel regression method, *Parity Regression (PR)*, and make three main contributions. *First,* we propose the PR estimator, which rather than minimising the aggregate prediction error, distributes the total prediction error evenly across all regression parameters, accompanied by a thorough theoretical characterisation. *Second,* we empirically show that our estimator outperforms OLS and some penalised and shrinkage regression methods on both synthetic and real datasets. *Third,* our proposed estimators are implemented in our **R** CRAN package *savvyPR*[1].

1.1 Literature review

This literature review begins with Stein's paradox (Stein, 1956; James and Stein, 1961), which marked a fundamental shift in statistical thinking by demonstrating that shrinkage can systematically improve estimation accuracy when measured by MSE. By showing that deliberate introduction of bias may reduce the overall estimation error through a variance–bias trade-off, it provided a conceptual foundation for modern regularisation techniques. Although shrinkage in its classical form emerged after Tikhonov regularisation (Tikhonov et al., 1943), which laid the groundwork for penalised regression, the underlying principle is

[1]Available at: https://github.com/Ziwei-ChenChen/savvyPR

closely related. The common thread linking penalised regression and shrinkage – despite their origins in different applications – is that a controlled introduction of bias can substantially reduce the estimator variability, thereby yielding an estimator that outperforms the natural unbiased estimator.

We begin by setting aside the estimation of the regression parameter vector $\boldsymbol{\theta}$ and instead consider the problem of estimating the population mean vector $\boldsymbol{\mu}$, thereby clarifying the foundations of the shrinkage principle arising from *Stein's paradox*. This paradox was introduced in the seminal papers (Stein, 1956; James and Stein, 1961) and has puzzled the statistical community ever since. The central message of the paradox is that unbiased estimators, even when endowed with strong theoretical and statistical properties, may nevertheless be "inefficient", where "*efficiency*" is measured in terms of MSE. Recall that the MSE of a generic estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is defined as

$$\mathrm{MSE}\big(\widehat{\boldsymbol{\theta}}\big) := \mathrm{Var}\big(\widehat{\boldsymbol{\theta}}\big) + \Big(\mathrm{Bias}\big(\widehat{\boldsymbol{\theta}}\big)\Big)^2. \tag{3}$$

The estimator proposed in (Stein, 1956; James and Stein, 1961), commonly referred to as the James–Stein estimator, demonstrates that the sample mean vector $\overline{\mathbf{X}} \in \mathbb{R}^p$ is a sub-optimal estimator of the population mean vector $\boldsymbol{\mu}$. Under the assumption of multivariate normal sampling distribution, a superior estimator in terms of MSE is obtained via multiplicative shrinkage, $\widehat{\boldsymbol{\mu}} = c\,\overline{\mathbf{X}}$, where c denotes the theoretical MSE-optimal shrinkage parameter. This estimator is often termed the *oracle* shrinkage estimator, as it still depends on unknown population parameters and is therefore not fully data-driven. A fully data-driven counterpart to the oracle estimator is known as the bona fide shrinkage estimator. For example, the James–Stein estimator derived in James and Stein (1961) is given by

$$\widehat{\boldsymbol{\mu}}_{\mathrm{PJS}} := \left(1 - \frac{(p-2)\hat{\sigma}^2}{n\|\overline{\mathbf{X}}\|_2^2}\right)_+ \overline{\mathbf{X}}, \qquad \text{for all } p \geq 3 \text{ and } n \geq 2,$$

where $t_+ := \max(t, 0)$ and $\hat{\sigma}^2 := \frac{1}{p}\,\mathrm{Tr}(\boldsymbol{S})$, with $\boldsymbol{S}$ denoting the sample covariance matrix estimator; note that $\|\cdot\|_p$ denotes the usual p-norm. For a comprehensive treatment of mean

vector shrinkage estimation, the reader is referred to (Bodnar et al., 2022; Asimit et al., 2026a).

Stein's paradox introduced not only mean vector estimators with low estimation error, but also a general shrinkage principle extending beyond the estimation of high-dimensional mean vectors. In particular, this principle can be applied to the estimation of the regression parameter vector $\boldsymbol{\theta}$. We therefore provide a succinct review of shrinkage estimators, which deliberately introduce bias in order to reduce the overall MSE through a variance–bias trade-off, a concept that lies at the core of *Cross Validation (CV)* in statistics and machine learning.

An alternative to OLS, introduced by Hoerl and Kennard (1970), is *Ridge Regression (RR)*, which is designed to mitigate overfitting by shrinking the regression parameters and is particularly useful in the presence of multicollinearity or ill-conditioning (when $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ has zero or very small eigenvalues). RR minimises the L_2-penalised $\mathcal{RSS}$ as follows:

$$\widehat{\boldsymbol{\theta}}^{RR}(\lambda) := \underset{\boldsymbol{\theta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\, \mathcal{RSS}(\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2^2, \qquad \lambda \geq 0, \tag{4}$$

where λ is a tuning parameter controlling the strength of the penalty. The solution admits the closed-form expression

$$\widehat{\boldsymbol{\theta}}^{RR}(\lambda) = \left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \lambda\mathbf{I}_{p+1}\right)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y}, \tag{5}$$

where $\lambda > 0$ guarantees that $\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \lambda\mathbf{I}_{p+1}\right)^{-1}$ exists. The penalisation term reduces estimation error, particularly when some eigenvalues of $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ are zero or close to zero.

By standard duality arguments, (4) is equivalent to the constrained formulation

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} \mathcal{RSS}(\boldsymbol{\theta}) \quad \text{subject to} \quad \sum_{k=0}^{p}\theta_k^2 \leq \tilde{\lambda}, \qquad \tilde{\lambda} \geq 0, \tag{6}$$

where $\tilde{\lambda}$ controls the size of the constraint set.

RR is a particular case of Tikhonov regularisation (Tikhonov et al., 1943), a broader framework for addressing ill-posed estimation problems. Specifically, for a penalty function

$g : \mathbb{R}^{p+1} \to \mathbb{R}_+$, the Tikhonov estimator is defined as

$$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + g(\boldsymbol{\theta}) \right\}. \tag{7}$$

When $g(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_2^2$, the Tikhonov estimator reduces to the RR estimator. Since

$$\widehat{\boldsymbol{\theta}}^{RR}(0) = \widehat{\boldsymbol{\theta}}^{OLS} \quad \text{and} \quad \widehat{\boldsymbol{\theta}}^{RR}(\lambda) \to \mathbf{0} \quad \text{as} \quad \lambda \to \infty,$$

RR is a *shrinkage estimator* that increasingly biases the estimates towards the origin as λ grows. Hoerl and Kennard (1970) showed that there exists an oracle estimator $\lambda^\star > 0$ such that

$$\text{MSE}\big(\widehat{\boldsymbol{\theta}}^{RR}(\lambda^\star)\big) < \text{MSE}\big(\widehat{\boldsymbol{\theta}}^{OLS}\big),$$

demonstrating that RR can outperform OLS when λ is suitably chosen. CV provides a practical method for selecting a bona fide estimate of $\lambda^\star$; however, the in-sample optimal choice may not be too close to the OOS optimal choice, which may increase the estimation error of the linear model.

When $g(\boldsymbol{\theta}) = t\|\boldsymbol{\theta}\|_1$ with $t \geq 0$, the Tikhonov estimator reduces to the *Least Absolute Shrinkage and Selection Operator (LASSO)* (Tibshirani, 1996) and *Basis Pursuit Denoising* (Chen and Donoho, 1994). The L_1-norm penalty promotes sparsity by shrinking some parameters exactly to zero, thereby performing variable selection in addition to regularisation. Owing to the non-differentiability of the L_1 penalty, LASSO does not admit a closed-form solution; however, it is equivalent to the constrained optimisation problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \mathcal{RSS}(\boldsymbol{\theta}) \quad \text{subject to} \quad \sum_{k=0}^{p} |\theta_k| \leq \tilde{t}, \qquad \tilde{t} \geq 0. \tag{8}$$

Although both LASSO and RR regularise the model, they differ fundamentally in their mechanisms. LASSO induces sparsity by setting certain parameters exactly to zero, thereby selecting a subset of predictors and enhancing interpretability. In contrast, RR shrinks the parameters continuously towards zero without eliminating any of them entirely.

RR and LASSO are examples of penalised regression methods that can be interpreted as

shrinkage estimators, although they are not constructed explicitly as shrinkage procedures. In contrast, there exists a broad class of estimators that directly shrink the OLS estimator towards a specified target. Such shrinkage estimators are typically simple, admit closed-form expressions, and are designed to optimise the theoretical MSE. Ideally, the corresponding oracle optimal shrinkage estimator is available in closed form, with its plug-in counterpart serving as a bona fide estimator, although Cross Validation may alternatively be employed. Both approaches present advantages and disadvantages at the implementation stage; however, this practical distinction has largely been overlooked in the existing literature.

The *Liu estimator (Liu)* (Liu, 1993) is a shrinkage estimator that directly shrinks the OLS estimator towards the target $\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}+\mathbf{I}_{p+1}\right)^{-1}\widetilde{\mathbf{X}}^\top\mathbf{y}$. Specifically, it modifies the OLS estimator as follows:

$$\widehat{\boldsymbol{\theta}}^{\text{Liu}}(d)=\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}+\mathbf{I}_{p+1}\right)^{-1}\left(\widetilde{\mathbf{X}}^\top\mathbf{y}+d\,\widehat{\boldsymbol{\theta}}^{\text{OLS}}\right), \tag{9}$$

where $d\in(0,1)$ is a shrinkage parameter. Under certain conditions, Liu (1993) showed that there exists an optimal $d^\star\in(0,1)$ such that

$$\text{MSE}\left(\widehat{\boldsymbol{\theta}}^{\text{Liu}}(d^\star)\right)<\text{MSE}\left(\widehat{\boldsymbol{\theta}}^{\text{OLS}}\right).$$

Hence, the oracle optimal shrinkage estimator $d^\star$ admits a closed-form expression. Nevertheless, in practice, standard software implementations typically select d via CV, similarly to the RR estimator. Finally, note that $\widehat{\boldsymbol{\theta}}^{\text{Liu}}(1)=\widehat{\boldsymbol{\theta}}^{\text{OLS}}$.

Liu (2003) extended this framework by proposing a *two-parameter Liu estimator* to address multicollinearity more effectively. The aim is for the Liu estimator to inherit the stabilising properties of the RR estimator, which is specifically designed to handle situations in which $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ possesses zero or near-zero eigenvalues, which is a hallmark of multicollinearity. The two-parameter Liu estimator introduces an additional parameter k to provide finer control over the shrinkage effect, while retaining the adjustment governed by d. It is defined

as

$$\widehat{\boldsymbol{\theta}}^{\text{Liu-type}}(k,d) = \big(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + k\mathbf{I}_{p+1}\big)^{-1}\big(\widetilde{\mathbf{X}}^\top\mathbf{y} - d\,\widehat{\boldsymbol{\theta}}^{RR}\big).$$

Liu (2003) showed that for any $k > 0$, there exists an optimal $d^{\star\star}$ such that

$$\text{MSE}\big(\widehat{\boldsymbol{\theta}}^{\text{Liu-type}}(k,d^{\star\star})\big) \leq \text{MSE}\big(\widehat{\boldsymbol{\theta}}^{RR}\big).$$

Although this estimator offers greater flexibility, both parameters must typically be selected via CV in practical implementations. Joint tuning of (k,d) is computationally expensive, and the additional estimation variability may increase the overall MSE, and thus, the two-parameter Liu estimator is less practical for real-world applications despite its attractive theoretical properties. For these reasons, we do not implement two-parameter Liu estimator in this paper.

The remainder of the paper is organised as follows. Section 2 presents the main theoretical results. Section 3 reports an informative simulation study, while Section 4 provides a comprehensive real-data analysis. Concluding remarks are given in Section 5. All proofs and supplementary material are collected in three appendices. Appendix A contains the proofs of all theoretical results. Appendix B provides additional details on the data-generating process underlying the simulation study. Appendix C includes further information on the datasets used in Section 4.

2 Main results

This section presents all the theoretical results of the paper. We begin with Section 2.1, which provides a general overview of parity estimation, including related theoretical results underpinning the concept. This foundational theory clarifies how parity estimation is particularly applied to multiple linear regression; for further details, see Section 2.2. The section concludes with Section 2.3, where we aim to enhance the explainability of the proposed concepts developed in Sections 2.1 and 2.2.

Before presenting the main results, we introduce some notation. The symbol $\succeq$ indicates that one symmetric matrix is greater than or equal to another in the Loewner ordering, meaning that their difference is positive semidefinite, whereas $\succ$ denotes strict dominance, meaning that their difference is positive definite. In particular, $\boldsymbol{\Sigma} \succ 0$ ($\boldsymbol{\Sigma} \succeq 0$) indicates that $\boldsymbol{\Sigma}$ is positive definite (positive semidefinite). Additionally, $\operatorname{diag}(\boldsymbol{A})$ denotes the diagonal matrix formed from the diagonal elements of the matrix $\boldsymbol{A}$.

2.1 Parity estimation

In statistics and machine learning, predictive models aim to relate a dependent target variable Y to a covariate (feature) vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Let $l : \mathbb{R} \times \mathbb{R}^p \times \Theta \to \mathbb{R}_+$ denote a loss function, where $\Theta \subset \mathbb{R}^q$ is the feasible set for the q-dimensional parameter vector $\boldsymbol{\theta}$. The model parameters $\boldsymbol{\theta}$ are estimated by minimising the expected loss, $\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}\big[l(Y, \mathbf{X}; \boldsymbol{\theta})\big]$. It is often assumed that $\mathcal{L} : \Theta \to (0, \infty]$ is differentiable, which we assume throughout the paper. The estimation problem therefore reduces to finding

$$\widehat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}). \tag{10}$$

We now introduce the notion of *parity estimation*, which, to the best of our knowledge, has not previously been studied. A parity estimator is a vector $\boldsymbol{\theta} \in \Theta$ such that $\theta_i \neq 0$ for all $i = 1, 2, \ldots, q$, and

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k} \Big/ \frac{\mathcal{L}(\boldsymbol{\theta})}{\theta_k} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_l} \Big/ \frac{\mathcal{L}(\boldsymbol{\theta})}{\theta_l} \quad \text{for all } 1 \leq k < l \leq q. \tag{11}$$

Condition (11) states that the elasticity of the loss function $\mathcal{L}$ is the same across all components of the parameter vector $\boldsymbol{\theta}$. This is inspired by the theoretical foundation of capital allocation in linear risk portfolios; for a brief overview of this concept, we refer the reader to (Tasche, 1999; Asimit et al., 2011, 2013, 2019). We are now ready to introduce the additional assumptions required for our main results, stated as Assumptions 1 and 2.

Assumption 1 *Assume that the feasible set $\Theta \subset \mathbb{R}^q$ is the parametric cone*

$$\mathcal{K}_q(\boldsymbol{\delta}) := \big\{\boldsymbol{\theta} \in \mathbb{R}^q : \delta_i \theta_i > 0,\ i = 1, \ldots, q\big\},$$

where $\boldsymbol{\delta} \in \{-1, 1\}^q$.

Assumption 2 *The loss function $\mathcal{L}$ satisfies the following growth condition: there exist constants $M_1, M_2 > 0$ such that*

$$\mathcal{L}(\boldsymbol{\theta}) \geq M_1 \|\boldsymbol{\theta}\|_\infty \quad \textit{for all } \boldsymbol{\theta} \in \Theta \textit{ with } \|\boldsymbol{\theta}\|_\infty > M_2, \textit{ where } \|\boldsymbol{\theta}\|_\infty = \max_i |\theta_i|.$$

We are now ready to present our first main result, stated as Theorem 1, which provides a characterisation of parity estimation.

Theorem 1 *Under Assumption 1, the following results hold.*

i) For any $\mu > 0$, any solution of (12)

$$\min_{\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta})} \Big(\mathcal{L}(\boldsymbol{\theta}) - \mu \sum_{k=1}^{q} \log(\delta_k \theta_k) \Big), \tag{12}$$

is a parity estimator.

ii) If $\mathcal{L}$ is convex and Assumption 2 holds, then (12) admits a unique solution for any $\mu > 0$.

iii) Assume that $\mathcal{L}(\boldsymbol{\theta})$ is convex and homogeneous of order $\tau \geq 1$. Then, for any $\mu > 0$, (12) admits a unique solution $\boldsymbol{\theta}^\star(\mu)$ which satisfies

$$\mathcal{L}(\boldsymbol{\theta}^\star(\mu)) = \frac{q\mu}{\tau} \quad \textit{and} \quad \boldsymbol{\theta}^\star(\mu) = \mu^{1/\tau} \boldsymbol{\theta}^\star(1).$$

Further, $\widetilde{\boldsymbol{\theta}} \in \mathcal{K}_q(\boldsymbol{\delta})$ is a parity estimator in (11) if and only if there exists $\mu_0 > 0$ such that $\widetilde{\boldsymbol{\theta}} = \mu_0^{1/\tau} \boldsymbol{\theta}^\star(1)$.

Condition (11) may be relaxed to define a *partial parity estimator*. Specifically, a vector

$\boldsymbol{\theta} \in \Theta$ is called a partial parity estimator if $\theta_i \neq 0$ for all $i = 1, 2, \ldots, q_0$, and

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_{k_1}} \Big/ \frac{\mathcal{L}(\boldsymbol{\theta})}{\theta_{k_1}} = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_{k_2}} \Big/ \frac{\mathcal{L}(\boldsymbol{\theta})}{\theta_{k_2}} \quad \text{for all } 1 \leq k_1 < k_2 \leq q_0. \tag{13}$$

Condition (13) states that the elasticity of the loss function $\mathcal{L}$ is identical across a specified subset of components of the parameter vector $\boldsymbol{\theta}$.

We are now ready to present our second main result, stated as Proposition 1, which provides a characterisation of partial parity estimation.

Proposition 1 *Under Assumption 1 and let* $\boldsymbol{t} = (t_{q_0+1}, t_{q_0+2}, \ldots, t_q) \geq \boldsymbol{0}_{q-q_0}$, *where* $\boldsymbol{0}_{q-q_0}$ *is a zero vector of dimension* $q - q_0$, *the following results hold.*

i) For any $\mu > 0$ *and* $\boldsymbol{t} \geq \boldsymbol{0}_{q-q_0}$, *any solution of* (14)

$$\min_{\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta})} \left(\mathcal{L}(\boldsymbol{\theta}) - \mu \sum_{k=1}^{q_0} \log(\delta_k \theta_k) - \mu \sum_{k=q_0+1}^{q} t_k \log(\delta_k \theta_k) \right), \tag{14}$$

is the parity estimator in (13).

ii) Assume that $\mathcal{L}$ *is convex and Assumption 2 holds. Then,* (14) *admits a unique solution for any* $\mu > 0$ *and* $\boldsymbol{t} \geq \boldsymbol{0}_{q-q_0}$.

iii) Assume that $\mathcal{L}(\boldsymbol{\theta})$ *is convex and homogeneous of order* $\tau \geq 1$ *in* $\boldsymbol{\theta}$, *then for any* $\mu > 0$ *and* $\boldsymbol{t} \geq \boldsymbol{0}_{q-q_0}$, (14) *admits a unique solution* $\boldsymbol{\theta}^\star(\mu, \boldsymbol{t})$ *which satisfies*

$$\mathcal{L}(\boldsymbol{\theta}^\star(\mu, \boldsymbol{t})) = \frac{(q_0 + \boldsymbol{1}^\top \boldsymbol{t})\mu}{\tau} \quad \text{and} \quad \boldsymbol{\theta}^\star(\mu, \boldsymbol{t}) = \mu^{1/\tau} \boldsymbol{\theta}^\star(1, \boldsymbol{t})$$

The proof of Proposition 1 is omitted, as it follows the same reasoning used in the proof of Theorem 1. Note that the final statement in Proposition 1 iii) does *not* hold on an "if and only if" basis, unlike its counterpart in Theorem 1 iii). Specifically, we cannot assert that any interior point $\widetilde{\boldsymbol{\theta}} \in \mathcal{K}_q(\boldsymbol{\delta})$ satisfying (13) necessarily implies the existence of $\mu_0 > 0$ and $\mathbf{t}_0 \geq \mathbf{0}_{q-q_0}$ such that $\widetilde{\boldsymbol{\theta}} = \mu_0^{1/\tau} \boldsymbol{\theta}^\star(1, \mathbf{t}_0)$. This means that the parametric optimal solutions in (14), with parameters $(\mu, \mathbf{t}) \in \mathbb{R}_{++} \times \mathbb{R}_+^{q-q_0}$, may yield a large set of vectors satisfying (13), but not necessarily all of them. This contrasts with Theorem 1, where there is a one-to-one

correspondence between the parametric set of optimal solutions in (12), parameterised by $\mu \in \mathbb{R}_{++}$, and the set of vectors satisfying (11). Consequently, it is practical to search for optimal solutions in (14) over $(\mu, \mathbf{0}_{q-q_0})$ with $\mu > 0$. That is, a partial parity estimator can be obtained as the parametric set of solutions in μ, given by

$$\min_{\boldsymbol{\theta} \in \mathcal{K}_{q_0}(\boldsymbol{\delta})} \mathcal{L}(\boldsymbol{\theta}) - \mu \sum_{k=1}^{q_0} \log(\delta_k \theta_k), \quad \text{with } \mu > 0, \tag{15}$$

since this formulation effectively minimises the loss with respect to the remaining components $(\theta_{q_0+1}, \ldots, \theta_q)$.

Section 2.2 provides a broad overview of how parity estimation operates in the context of multiple linear regression, while also presenting some specific theoretical results for linear regression models.

2.2 Parity estimation for linear regression

We begin by noting that the *Parity Regression (PR)* – the parity estimation framework for linear regression – shares conceptual similarities with well-known penalised regression methods, as it introduces specific constraints. Figure 1 provides a simple geometric interpretation of PR estimation, reminiscent of the classic geometric representations of RR; for example, see Figure 3.11 in Hastie et al. (2009) or Figure 2 in Tibshirani (1996).

The illustration in Figure 1 provides an intuitive explanation for the presence of a single solution in each quadrant when $p = 1$, and the same reasoning extends naturally to cases with $p > 1$. Theorem 1 and Proposition 1 provide the theoretical justification for this geometric interpretation.

The PR method builds upon the structure of RR by introducing elasticity constraints that ensure the resulting loss function is homogeneous and convex in a specific form. This can be formalised by defining the total expected loss as

$$\mathcal{RRSS}(\widehat{\boldsymbol{\theta}}; \lambda) \;:=\; \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \tilde{\mathbf{x}}_i - \theta_{p+1} y_i \right)^2 + \lambda \sum_{k=1}^{p} \theta_k^2, \tag{16}$$

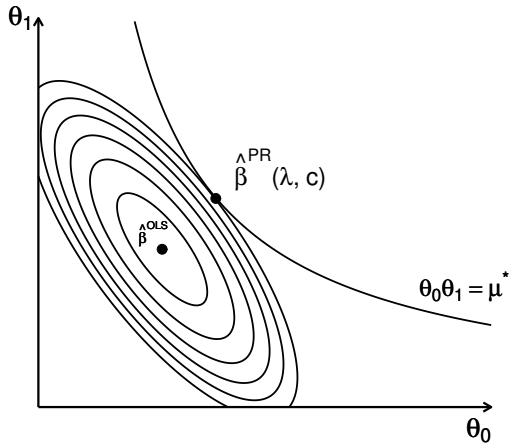


Figure 1: Geometric interpretation of PR estimation in $\mathcal{K}_2(1,1)$ for $p = 1$, $\lambda = 0$, and an appropriately chosen $\mu^\star > 0$.

$$= \frac{1}{n}\widehat{\boldsymbol{\theta}}^\top \mathbf{Z}^\top \mathbf{Z}\widehat{\boldsymbol{\theta}} + \lambda \widehat{\boldsymbol{\theta}}^\top \operatorname{diag}(\mathbf{1}_{p+1}, 0)\boldsymbol{\theta},$$

where $\widehat{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \theta_{p+1})$ extends the parameter vector by including the synthetic term $\theta_{p+1} = 1$ for the dependent variable, $\mathbf{1}_{p+1}$ is a vector of ones of length $p+1$, and $\mathbf{Z}$ is an $n \times (p+2)$ matrix with i^{th} row given by $\left(\tilde{\mathbf{x}}_i^\top, -y_i\right)$ for all $1 \leq i \leq n$. To satisfy the homogeneity condition required by the parity estimator, as stated in Proposition 1, we introduce the synthetic term $\theta_{p+1} = 1$. This ensures that (16) becomes a homogeneous function of order $\tau = 2$, while maintaining convexity and homogeneity for the extended parameter vector $\widehat{\boldsymbol{\theta}} \in \mathcal{K}_{p+2}(\boldsymbol{\delta})$. These properties fulfill the necessary conditions for applying parity estimation, thereby allowing us to directly extend parity estimation principles to the PR framework. To ensure that the loss function $\mathcal{RRSS}$ behaves as intended, we impose the following conditions, stated formally in Assumption 3.

Assumption 3 *Let Assumption 1 hold. If $\lambda = 0$, then*

$$\mathcal{RRSS}\big((\boldsymbol{\theta}, 1); 0\big) > 0 \quad \textit{for all } (\boldsymbol{\theta}, 1) \in \mathcal{K}_{p+2}(\boldsymbol{\delta}). \tag{17}$$

Note that $\mathbf{Z}^\top \mathbf{Z} \succeq 0$, and therefore, for any given $\lambda > 0$ and search cone $\mathcal{K}_{p+2}(\boldsymbol{\delta})$, we have

$$\mathcal{RRSS}\big((\boldsymbol{\theta}, 1); \lambda\big) > 0 \quad \text{for all } (\boldsymbol{\theta}, 1) \in \mathcal{K}_{p+2}(\boldsymbol{\delta}). \tag{18}$$

This condition may fail only if there exists $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$ such that

$$y_i - \tilde{\theta}_0 - \sum_{k=1}^{p} \tilde{\theta}_k x_{ik} = 0 \quad \text{for all } 1 \leq i \leq n, \tag{19}$$

with reference to (1). Specifically, (19) can arise in the following scenarios: (i) an imbalanced regression with $n < p+1$, (ii) strong linear dependence among some features/predictors with $n > p+1$, or (iii) a special data structure when $n = p+1$ such that $\widehat{\boldsymbol{\theta}}^{OLS} = \widetilde{\boldsymbol{\theta}}$. In these cases, Assumption 3 implies

$$\mathcal{RRSS}\big((\boldsymbol{\theta}, 1); 0\big) = 0 \quad \text{if and only if } \boldsymbol{\theta} \in \mathbb{R}^{p+1} \setminus \{\widetilde{\boldsymbol{\theta}}\}.$$

Our main PR results for this section, namely Theorem 2 and Proposition 2, are now presented. These results establish the existence and uniqueness of the PR estimators under the specified elasticity-based constraints.

Theorem 2 *Let $\lambda \geq 0$ and $t \geq 0$ for which Assumptions 1 and 3 hold.*

i) For any $\tilde{\mu} > 0$, the unconstrained optimisation problem

$$\min_{(\boldsymbol{\theta}, \theta_{p+1}) \in \mathcal{K}_{p+2}(\boldsymbol{\delta})} \left(\mathcal{RRSS}(\boldsymbol{\theta}, \theta_{p+1}; \lambda) - \mu \sum_{k=0}^{p} \log(\delta_k \theta_k) - \mu t \log(\delta_{p+1} \theta_{p+1}) \right) \tag{20}$$

admits a unique solution, denoted as $\big((\boldsymbol{\theta}^\star(\lambda, t, \mu))^\top, \theta^\star_{p+1}(\lambda, t, \mu)\big)^\top$, which satisfies the parity conditions in (11). This solution yields the PR estimate

$$\widehat{\boldsymbol{\theta}}^{PR}(\lambda, t) = \frac{\big((\boldsymbol{\theta}^\star(\lambda, t, \mu))^\top, \theta^\star_{p+1}(\lambda, t, \mu)\big)^\top}{\theta^\star_{p+1}(\lambda, t, \mu)} = \big((\widehat{\boldsymbol{\beta}}^{PR}(\lambda, t))^\top, 1\big)^\top, \tag{21}$$

which is independent of $\mu > 0$ for any fixed (λ, t). Furthermore, we have

$$\mathcal{RRSS}\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, \mu), \theta^\star_{p+1}(\lambda, t, \mu)\big); \lambda\Big) = \frac{(p+1+t)\mu}{2}, \tag{22}$$

for any $(\lambda, t, \mu) \in \mathbb{R}_+ \times [0, \infty) \times \mathbb{R}^\star_+$, and

$$\widehat{\boldsymbol{\beta}}^{PR}(\lambda, t) = \boldsymbol{\theta}^\star(\lambda, t, \mu^\star) = (\mu^\star)^{1/2} \boldsymbol{\theta}^\star(\lambda, t, 1), \tag{23}$$

where $\mu^\star = \big(\theta^\star_{p+1}(\lambda,t,1)\big)^{-2}$. *Furthermore, define* $\widetilde{\boldsymbol{\theta}}^{PR}(\lambda,t) := \big(\big(\widetilde{\boldsymbol{\beta}}^{PR}(\lambda,t)\big)^\top, 1\big)^\top$ *as a PR estimate in the search cone* $\mathcal{K}_{p+2}(\boldsymbol{\delta})$, *which satisfies* (13). *Then,* $\widetilde{\boldsymbol{\theta}}^{PR}(\lambda,t)$ *is the unique solution of* (20) *with* $\tilde{\mu} = \Big(\frac{2}{p+1+t}\Big)\mathcal{RRSS}\Big(\widetilde{\boldsymbol{\beta}}^{PR}(\lambda,t), 1; \lambda\Big)$. *Consequently, we have* $\widehat{\boldsymbol{\beta}}^{PR}(\lambda,t) = \widetilde{\boldsymbol{\beta}}^{PR}(\lambda,t)$.

ii) For any $\tilde{\mu} \in \mathbb{R}$, *the constrained optimisation problem*

$$\begin{cases} \min\limits_{(\boldsymbol{\theta},\theta_{p+1})\in\mathcal{K}_{p+2}(\boldsymbol{\delta})} & \mathcal{RRSS}\big(\boldsymbol{\theta},\theta_{p+1};\lambda\big) \\ s.t. & \sum\limits_{k=0}^{p}\log(\delta_k\theta_k) + t\log(\delta_{p+1}\theta_{p+1}) \geq \tilde{\mu} \end{cases} \tag{24}$$

admits a unique solution, denoted as $\big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top, \theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})\big)^\top$. *This solution yields the PR estimate*

$$\widehat{\widetilde{\boldsymbol{\theta}}}^{PR}(\lambda,t) = \frac{\big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top, \theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})\big)^\top}{\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})} = \left(\left(\widehat{\widetilde{\boldsymbol{\beta}}}^{PR}(\lambda,t)\right)^\top, 1\right)^\top, \tag{25}$$

which is independent of $\tilde{\mu}$ *for any given* (λ,t). *Additionally, we have*

$$\widehat{\widetilde{\boldsymbol{\beta}}}^{PR}(\lambda,t) = \boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}^\star) = e^{\frac{\tilde{\mu}^\star}{p+1+t}}\boldsymbol{\theta}^{\star\star}(\lambda,t,0), \tag{26}$$

where $\tilde{\mu}^\star = -(p+1+t)\log\theta^{\star\star}_{p+1}(\lambda,t,0)$, *and strong duality holds in* (24).

iii) For any given values of $\mu > 0$ *and* $\tilde{\mu} \in \mathbb{R}$, *we have that* $\widehat{\boldsymbol{\beta}}^{PR}(\lambda,t) = \widehat{\widetilde{\boldsymbol{\beta}}}^{PR}(\lambda,t)$.

Theorem 2 i) shows that up to 2^{p+1} distinct PR estimates can be identified, each residing in one of the 2^{p+1} possible search cones. This is consistent with Assumption 1, where estimators are sought within specific quadrants of the parameter space. For a given p, the PR estimates can therefore span all possible quadrants, and (20) provides a systematic procedure to obtain each potential PR solution in this setting.

Theorem 2 ii) represents a constrained form of i) and constitutes a special case of the framework discussed in Section 2.3. Theorem 2 iii) establishes the uniqueness of each PR estimate and its independence from the normalising constants μ and $\tilde{\mu}$, thereby simplifying computation by eliminating the need for cross-validation over these constants. As a result,

any suitable computational method, such as (20), produces a stable, single PR estimate within the chosen cone $\mathcal{K}_{p+2}(\boldsymbol{\delta})$ across all parameter quadrants when $p > 1$.

We are now ready to present the main result of this section, stated as Proposition 2.

Proposition 2 *Assume that $\widehat{\boldsymbol{\beta}}^{RR}(\lambda)$ and $\widehat{\boldsymbol{\beta}}^{OLS}$ contain only non-zero elements. Let*

$$\boldsymbol{\delta}^{RR} := sgn\Big(\widehat{\boldsymbol{\beta}}^{RR}(\lambda)\Big) \quad \text{and} \quad \boldsymbol{\delta}^{OLS} := sgn\Big(\widehat{\boldsymbol{\beta}}^{OLS}\Big),$$

where the signum function is applied componentwise such that $sgn(a) = 1$ and $sgn(a) = -1$ whenever $a > 0$ and $a < 0$, respectively. Thus, $\widehat{\boldsymbol{\beta}}^{RR}(\lambda) \in \mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{RR}\big)$ and $\widehat{\boldsymbol{\beta}}^{OLS} \in \mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{OLS}\big)$.

i) Let $\lambda > 0$ and $t \geq 0$ for which Assumption 1 holds for $\mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{RR}\big)$. The PR estimate $\widehat{\boldsymbol{\beta}}^{PR}(\lambda, c) \in \mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{RR}\big)$ satisfies

$$\prod_{k=0}^{p} \frac{\widehat{\boldsymbol{\beta}}_k^{PR}(\lambda, t)}{\widehat{\boldsymbol{\beta}}_k^{RR}(\lambda)} \geq 1. \tag{27}$$

ii) Let $\lambda = 0$ and $t \geq 0$ for which Assumption 1 holds for $\mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{OLS}\big)$. The PR estimate $\widehat{\boldsymbol{\beta}}^{PR}(0, t) \in \mathcal{K}_{p+1}\big(\boldsymbol{\delta}^{OLS}\big)$ satisfies

$$\prod_{k=0}^{p} \frac{\widehat{\boldsymbol{\beta}}_k^{PR}(0, t)}{\widehat{\boldsymbol{\beta}}_k^{OLS}} \geq 1. \tag{28}$$

Proposition 2 recommends using either the OLS or RR estimate as a starting point for selecting the search cone in PR estimation. Choosing the cone defined by OLS, $\widetilde{\boldsymbol{\theta}}^{OLS} = \big(\widehat{\boldsymbol{\beta}}^{OLS}, 1\big) = \big(\widehat{\boldsymbol{\beta}}^{RR}(0), 1\big)$, provides a non-regularised baseline. Alternatively, selecting the cone containing the RR estimate, $\widetilde{\boldsymbol{\theta}}^{RR} = \big(\widehat{\boldsymbol{\beta}}^{RR}(\lambda^\star), 1\big)$, offers a regularised approach, where $\lambda^\star$ is chosen via CV, for example using the *glmnet* package in **R**.

2.3 Parity estimation and regression – further explainability

We have established the main theory of parity estimation and PR in the previous two sections. We now aim to enhance the explainability of these concepts by providing a higher-level

description of the theory and by linking PR to well-known penalised regression methods, such as RR and LASSO. To this end, we introduce the concept of the *Generalised Weighted Mean (GWM)*.

The GWM generalises several well-known averaging operations and depends on a parameter r. For a given $\mathbf{x} \in \mathbb{R}^{p+1}$ and a vector of weights $\boldsymbol{b} = (b_0, b_1, \ldots, b_p)^\top$, where $b_i \geq 0$ and $\mathbf{1}^\top \boldsymbol{b} = 1$, the GWM of order r is defined as

$$m_r(\mathbf{x}; \boldsymbol{b}) = \left(\sum_{k=0}^{p} b_k |x_k|^r \right)^{\frac{1}{r}}, \quad \text{for } r \in \mathbb{R} \cup \{\pm\infty\}.$$

It holds that $m_r(\mathbf{x}; \boldsymbol{b}) \leq m_s(\mathbf{x}; \boldsymbol{b})$ for all $-\infty \leq r < s \leq \infty$. The limiting case $r = 0$, corresponding to the weighted geometric mean, is given by

$$m_0(\mathbf{x}; \boldsymbol{b}) := \lim_{r \to 0} m_r(\mathbf{x}; \boldsymbol{b}) = \prod_{k=0}^{p} |x_k|^{b_k} = \exp\left(\sum_{k=0}^{p} b_k \log |x_k| \right).$$

For $r = \pm\infty$, we obtain

$$m_{-\infty}(\mathbf{x}; \boldsymbol{b}) = \min_{0 \leq i \leq p} |x_i| \qquad \text{and} \qquad m_{\infty}(\mathbf{x}; \boldsymbol{b}) = \max_{0 \leq i \leq p} |x_i|.$$

The case $r = -1$ yields the weighted harmonic mean, defined as

$$m_{-1}(\mathbf{x}; \boldsymbol{b}) = \left(\sum_{i=0}^{p} \frac{b_i}{|x_i|} \right)^{-1}.$$

We introduce the *Generalised Weighted Mean Constrained (GWMC)* estimation framework and note that an equivalent formulation in the context of portfolio theory is discussed in (Asimit et al., 2025, 2026c). The GWMC approach seeks to minimise a given loss function $\mathcal{L}$ subject to a constraint on the GWM of the regression parameters:

$$\begin{cases} \min\limits_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} & \mathcal{L}(\boldsymbol{\theta}) \\ \text{s.t.} & m_r(\boldsymbol{\theta}; \boldsymbol{b}) \leq \epsilon, \quad \text{for } r \geq 1, \\ & m_r(\boldsymbol{\theta}; \boldsymbol{b}) \geq \epsilon, \quad \text{for } r < 1, \end{cases} \tag{29}$$

where $\epsilon > 0$ is a fixed constant. The GWM function $m_r(\boldsymbol{\theta}; \boldsymbol{b})$ provides flexibility in regular-

ising the parameters through the choice of the order r and weighting vector $\boldsymbol{b}$.

Note that $m_r(\boldsymbol{\theta};\boldsymbol{b})$ is convex in $\boldsymbol{\theta} \in \mathcal{K}_{p+1}(\boldsymbol{\delta})$ when $r \geq 1$, and therefore encompasses a broad class of regularisation schemes. In particular, RR and LASSO arise as special cases of (29) when the loss functional $\mathcal{L}$ corresponds to a regression setting. Specifically, RR can be formulated within the GWMC framework with $r = 2$ and $\boldsymbol{\theta} \in \mathcal{K}_{p+1}(\boldsymbol{\delta})$:

$$m_2(\boldsymbol{\theta};\boldsymbol{b}) = \left(\sum_{k=0}^{p} b_k|\theta_k|^2\right)^{\frac{1}{2}} \leq \sqrt{\tilde{\lambda}},$$

where equal weights are assumed. This formulation coincides with the usual L_2-norm constraint in (6). Similarly, LASSO is representable within the GWMC framework with $r = 1$, which imposes an L_1-norm constraint on $\boldsymbol{\theta} \in \mathcal{K}_{p+1}(\boldsymbol{\delta})$. In this case, the GWMC constraint becomes

$$m_1(\boldsymbol{\theta};\boldsymbol{b}) = \sum_{k=0}^{p} b_k|\theta_k| \leq \tilde{t}.$$

The *Parity Estimator* introduced in (11) and formulated in (12) can be viewed as a limiting case of the GWMC framework in (29) with $r = 0$ and $\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta})$, which leads to a logarithmic constraint. Specifically, for equal weights $b_k = \frac{1}{q}$, the constraint becomes

$$\begin{aligned} m_0(\boldsymbol{\theta};\boldsymbol{b}) &:= \lim_{r\to 0} m_r(\boldsymbol{\theta};\boldsymbol{b}) \\ &= \exp\left(\sum_{i=1}^{q} b_i \log|\theta_i|\right) \geq e^{\mu}, \end{aligned}$$

where μ is a lower-bound parameter. Similarly, the Partial Parity Estimator defined in (13) and reformulated in (14) can also be interpreted within the GWMC framework with $r = 0$ and $\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta})$. In this case, the logarithmic constraint takes the weighted form

$$m_0(\boldsymbol{\theta};\boldsymbol{b}) := \lim_{r\to 0} m_r(\boldsymbol{\theta};\boldsymbol{b}) = \exp\left(\sum_{i=1}^{q_0} b_i \log|\theta_i| + \sum_{i=q_0+1}^{q} b_i \log|\theta_i|\right) \geq e^{\tilde{\mu}},$$

where $\tilde{\mu}$ is the corresponding lower-bound parameter. The weights b_k are defined as

$$b_k = \begin{cases} \dfrac{1}{q_0 + \mathbf{1}^\top \mathbf{t}}, & \text{for } k = 1, \ldots, q_0, \\ \dfrac{t_k}{q_0 + \mathbf{1}^\top \mathbf{t}}, & \text{for } k = q_0 + 1, \ldots, q. \end{cases}$$

In summary, the logarithmic constraints underlying parity estimation induce a balanced regularisation effect, distributing the elasticity of the loss function evenly across the selected parameters. When embedded within the GWMC framework, the parity estimator becomes conceptually aligned with penalised regression methods such as RR and LASSO, which correspond to the cases $r = 2$ and $r = 1$, respectively. The fundamental similarity amongst RR, LASSO, and PR lies in their shared objective: controlling both the magnitude and the distribution of model parameters through constraints on a weighted mean functional. This connection highlights how PR extends classical regularisation principles by incorporating fairness-oriented constraints within the unified GWMC framework.

3 Simulation study

In this section, we conduct a simulation study to evaluate the finite-sample performance of the PR estimators relative to their primary competitors: (i) OLS as defined in (2), (ii) RR as defined in (5), and (iii) the Liu estimator as defined in (9). Our objective is to assess the robustness of the PR framework under varying degrees of multicollinearity and heteroscedasticity. We exclude the LASSO estimator from the analysis, as our focus is on dense parameter space structures and variable selection is not the purpose of this paper.

This section is organised into two parts. Section 3.1 describes the simulation design and data-generating process, but it also introduces the performance measures used to compare the estimators. Section 3.2 presents and discusses the simulation results.

3.1 Experimental setup and methodology

To assess model performance across a broad spectrum of data structures, we utilise a *Data Generation Process (DGP)* designed to simulate complex regression environments characterised by both high multicollinearity and heteroscedasticity. A detailed characterisation of the simulation steps, including the construction of the feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the heteroscedastic response variable Y, is provided in Appendix B.

We compare several distinct estimation methodologies. The OLS estimator serves as the unregularised baseline, while the shrinkage estimators, namely RR and Liu, are also included. The PR framework is implemented in two computationally feasible versions, distinguished by how their parameters are tuned. *First,* the PR estimator with t-tuning, denoted as PR_t, is directly motivated by Theorem 2. Its key feature is that, for any fixed $t \geq 0$, the PR estimate is independent of the normalising constant μ. Here, t acts as the relative elasticity weight for the target variable and is selected via CV. *Second,* the PR estimator with c-tuning, denoted as PR_c, allocates a fixed loss contribution to each predictor. Mathematically, this corresponds to the budget-based objective function, where the logarithmic barrier for the p predictors is weighted by c, while the response variable's contribution is determined by $1 - (p+1)c$. To ensure positive risk allocations, c is constrained such that $0 \leq c < 1/(p+1)$. While c and t are functionally related, as both govern the balance of loss contributions, PR_c operates within a strictly bounded interval, offering an alternative computational perspective on the risk-balancing mechanism. Both PR_t and PR_c are compared against the three benchmarks (OLS, RR, and Liu), with the search cone for PR regressions determined by the sign of the OLS estimates, as suggested by Proposition 2. All tuning parameters (c, t, λ, or d) are selected via 10-fold CV.

To evaluate the accuracy of the proposed estimators in the simulation study, we measure their performance using the L_2-distance between the true regression parameter vector $\boldsymbol{\beta}$ and the estimated vector $\hat{\boldsymbol{\beta}}$. This corresponds to the estimated MSE of $\boldsymbol{\beta}$, which is appropriate

since low estimation error aligns with low theoretical MSE. For a single simulation run, the L_2-error is defined as

$$L_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{m} (\hat{\beta}_j - \beta_j)^2}, \tag{30}$$

where m denotes the number of covariates, including the intercept.

We report the average L_2-distance across $N = 1,000$ repetitions for each scenario. To assess the precision of these averages, we also report the *standard error (SE)*. Following the convention in our numerical results, the standard error is expressed as a percentage to facilitate comparison of the performance of different estimators and is calculated as follows:

$$\text{SE} = \left(\frac{\text{sd}(L_2)}{\sqrt{N}} \right) \times 100, \tag{31}$$

where $\text{sd}(L_2)$ denotes the standard deviation of the L_2-distances across all N repetitions. A lower average L_2-distance indicates greater accuracy in estimating the true regression parameters.

3.2 Discussion of simulation results

We provide a comprehensive discussion of the comparative performance of the two PR variants (PR_c and PR_t) relative to the OLS, RR, and the Liu estimators in Tables 1 and 2. The detailed results are reported in Table 1, while the aggregated results are summarised in Table 2 to facilitate identification of the main trends observed in Table 1.

Overall, the simulation results demonstrate that PR consistently outperforms OLS and the shrinkage estimators by effectively stabilising parameter estimates in high-correlation environments. While OLS exhibits substantial variance inflation as $|\rho|$ increases, both PR_c and PR_t achieve their strongest performance under high negative correlation ($\rho = -0.75$ and -0.5), frequently attaining the lowest L_2-distances across all panels. Even in settings with strong positive correlation, PR remains superior or highly competitive relative to OLS. Moreover, as dimensionality increases (Panel C, $m = 25$), PR_c emerges as the dominant

estimator across nearly all correlation levels. These findings confirm that the parity constraint provides an effective regularisation mechanism for high-dimensional models characterised by dense parameter structures and strong interdependence among features.

Table 2 provides a synthesised overview of the results reported in Table 1 by aggregating the "best" and "second-best" performances across all 60 scenarios. The summary highlights the consistency of the PR framework: the combined PR approach is the top-performing model in 41.7% of cases and ranks as either the best or second-best estimator in 86.7% of the scenarios considered (52 out of 60).

In summary, the simulation results indicate that PR estimators are highly competitive relative to all benchmarks, with particularly strong performance in settings characterised by pronounced multicollinearity and high dimensionality. By imposing structural balance on the parameters through parity constraints, the proposed methods achieve a more favourable bias–variance trade-off than the shrinkage estimators across a wide range of dependence structures.

4 Real data analysis

We evaluate the performance of the PR estimators (PR_c, PR_t) using data from the *West Texas Intermediate (WTI)* and *Brent* crude oil markets. These commodities are selected as they serve as the primary global benchmarks for oil pricing and are characterised by high volatility and frequent structural shifts. This setting enables us to assess the performance of the parity framework relative to OLS, RR, and the Liu estimator across diverse market regimes. Section 4.1 describes the dataset and the underlying factor model. The empirical results are presented and discussed in Section 4.2, but we also outline the OOS evaluation framework and the error metrics employed.

Table 1: **Simulation study results for different numbers of covariates**

	$n/m = 10$					$n/m = 25$					$n/m = 50$					$n/m = 100$				
ρ	-0.75	-0.5	0	0.5	0.75	-0.75	-0.5	0	0.5	0.75	-0.75	-0.5	0	0.5	0.75	-0.75	-0.5	0	0.5	0.75
Panel A: $m = 2$																				
OLS	4.15	3.23	2.82	3.23	4.15	2.66	2.07	1.80	2.07	2.66	1.81	1.43	1.27	1.43	1.81	1.31	1.02	0.90	1.02	1.31
	(8.98)	(6.21)	(4.85)	(6.21)	(8.98)	(5.26)	(3.69)	(2.97)	(3.69)	(5.26)	(3.59)	(2.54)	(2.12)	(2.54)	(3.59)	(2.49)	(1.77)	(1.47)	(1.77)	(2.49)
RR	2.03	**1.78**	**1.70**	**1.91**	**2.23**	1.47	1.34	1.36	1.55	1.77	1.14	1.09	1.17	1.35	1.52	0.87	0.85	0.93	1.13	1.34
	(7.16)	(4.71)	(3.34)	(4.47)	(6.77)	(4.04)	(2.63)	(1.82)	(2.16)	(3.35)	(2.62)	(1.81)	(1.45)	(1.39)	(1.85)	(1.81)	(1.36)	(1.24)	(1.30)	(1.31)
Liu	**2.02**	1.79	1.73	1.95	2.29	**1.38**	**1.24**	**1.25**	**1.45**	**1.72**	**1.08**	**1.01**	**1.06**	1.22	1.41	0.91	0.83	0.85	1.00	1.18
	(6.66)	(4.59)	(3.58)	(4.40)	(6.21)	(4.04)	(2.81)	(2.19)	(2.54)	(3.61)	(2.61)	(1.86)	(1.54)	(1.66)	(2.13)	(1.74)	(1.27)	(1.15)	(1.29)	(1.52)
PR_c	2.17	1.89	1.77	1.94	2.25	1.58	1.40	1.36	1.51	1.75	1.17	1.08	1.13	1.31	1.49	**0.85**	**0.82**	0.91	1.13	1.34
	(6.99)	(4.65)	(3.52)	(4.51)	(6.66)	(4.04)	(2.73)	(2.06)	(2.49)	(3.55)	(2.66)	(1.97)	(1.69)	(1.62)	(2.08)	(1.77)	(1.43)	(1.30)	(1.30)	(1.27)
PR_t	2.30	1.93	1.77	1.92	2.24	1.75	1.46	1.35	1.47	1.72	1.36	1.13	1.09	**1.20**	**1.40**	1.05	0.83	**0.77**	**0.92**	**1.13**
	(6.96)	(4.67)	(3.54)	(4.61)	(6.73)	(4.10)	(2.86)	(2.13)	(2.59)	(3.70)	(3.03)	(2.25)	(1.82)	(1.93)	(2.46)	(2.72)	(1.97)	(1.58)	(1.64)	(1.94)
Panel B: $m = 10$																				
OLS	5.83	4.06	3.24	**4.05**	**5.82**	3.60	2.51	1.99	**2.51**	**3.60**	2.56	1.78	1.42	**1.78**	**2.56**	1.79	1.25	**0.99**	**1.25**	**1.79**
	(5.20)	(3.39)	(2.42)	(3.36)	(5.16)	(3.18)	(2.07)	(1.46)	(2.08)	(3.19)	(2.28)	(1.48)	(1.03)	(1.48)	(2.28)	(1.59)	(1.04)	(0.72)	(1.04)	(1.59)
RR	**3.15**	**3.09**	3.15	4.23	6.38	**2.35**	2.16	1.98	2.54	3.70	1.88	1.63	1.41	1.80	2.60	1.44	1.18	0.99	1.25	1.81
	(3.26)	(2.51)	(2.38)	(3.94)	(6.47)	(2.23)	(1.72)	(1.45)	(2.19)	(3.49)	(1.75)	(1.32)	(1.04)	(1.50)	(2.36)	(1.27)	(0.96)	(0.73)	(1.05)	(1.65)
Liu	4.27	3.49	3.16	4.16	5.87	3.17	2.36	1.98	2.55	3.69	2.40	1.73	1.41	1.80	2.60	1.74	1.23	0.99	1.26	1.81
	(4.04)	(2.82)	(2.41)	(3.69)	(5.36)	(2.81)	(1.92)	(1.44)	(2.22)	(3.43)	(2.13)	(1.42)	(1.04)	(1.51)	(2.39)	(1.53)	(1.01)	(0.72)	(1.07)	(1.67)
PR_c	3.44	3.15	3.12	4.20	6.31	2.41	2.09	**1.95**	2.55	3.72	**1.83**	1.54	**1.38**	1.81	2.62	1.35	1.11	0.99	1.25	1.80
	(3.77)	(2.78)	(2.40)	(3.92)	(6.47)	(2.57)	(1.82)	(1.46)	(2.15)	(3.41)	(1.92)	(1.32)	(1.04)	(1.51)	(2.37)	(1.25)	(0.92)	(0.71)	(1.05)	(1.65)
PR_t	4.19	3.26	**3.10**	4.21	6.18	2.68	**2.07**	1.99	2.78	4.01	1.93	**1.52**	1.51	2.19	3.20	**1.33**	**1.10**	1.22	1.90	2.76
	(5.03)	(3.34)	(2.52)	(4.06)	(6.16)	(3.82)	(2.38)	(1.61)	(2.19)	(3.53)	(3.01)	(1.77)	(1.07)	(1.52)	(2.28)	(1.95)	(0.91)	(0.66)	(0.97)	(1.32)
Panel C: $m = 25$																				
OLS	6.17	4.25	**3.31**	**4.25**	**6.18**	3.75	2.58	2.03	**2.58**	**3.75**	2.63	1.81	1.42	**1.81**	**2.63**	1.85	1.27	1.00	**1.27**	1.85
	(3.53)	(2.30)	(1.56)	(2.29)	(3.54)	(2.02)	(1.32)	(0.91)	(1.31)	(2.03)	(1.44)	(0.93)	(0.64)	(0.93)	(1.43)	(1.01)	(0.66)	(0.46)	(0.66)	(1.01)
RR	**4.66**	3.99	3.31	4.26	6.18	3.22	2.52	2.03	2.59	3.75	2.41	1.79	1.42	1.81	2.63	1.75	1.27	**0.99**	1.27	1.85
	(2.68)	(2.08)	(1.58)	(2.30)	(3.65)	(1.73)	(1.28)	(0.91)	(1.31)	(2.05)	(1.30)	(0.91)	(0.64)	(0.93)	(1.44)	(0.95)	(0.66)	(0.45)	(0.66)	(1.01)
Liu	5.97	4.19	3.31	4.26	6.19	3.70	2.58	2.03	2.60	3.75	2.63	1.81	1.42	1.81	2.66	1.85	1.27	1.00	1.27	1.85
	(3.40)	(2.25)	(1.57)	(2.35)	(3.60)	(1.99)	(1.31)	(0.91)	(1.34)	(2.02)	(1.44)	(0.93)	(0.64)	(0.93)	(1.46)	(1.01)	(0.66)	(0.46)	(0.66)	(1.01)
PR_c	4.77	**3.89**	3.31	4.29	6.20	**3.08**	**2.43**	**2.02**	2.59	3.78	**2.27**	**1.74**	**1.41**	1.82	2.64	1.68	**1.24**	1.00	1.27	**1.84**
	(3.37)	(2.18)	(1.59)	(2.29)	(3.61)	(1.86)	(1.27)	(0.91)	(1.32)	(2.05)	(1.29)	(0.90)	(0.64)	(0.94)	(1.45)	(0.93)	(0.64)	(0.45)	(0.66)	(1.01)
PR_t	5.16	3.96	3.43	4.76	7.16	3.22	2.43	2.23	3.43	5.45	2.30	1.75	1.73	2.96	4.89	**1.66**	1.31	1.44	2.70	4.60
	(4.72)	(2.68)	(1.71)	(2.48)	(3.59)	(2.98)	(1.50)	(1.01)	(1.46)	(2.11)	(1.94)	(0.92)	(0.70)	(1.13)	(1.56)	(0.97)	(0.58)	(0.49)	(0.84)	(1.09)

Notes. This table reports the average L_2-distances, as defined in (30), with the corresponding standard errors (in brackets) defined in (31), based on $N = 1,000$ independent repetitions with sample size $n = 1,000$. We compare OLS, RR, Liu, and the two PR variants (PR_c, PR_t) across correlation parameters $\rho \in \{-0.75, -0.5, 0, 0.5, 0.75\}$ and ratios $n/m \in \{10, 25, 50, 100\}$. Results are organised into Panel A ($m = 2$), Panel B ($m = 10$), and Panel C ($m = 25$). **Red** values indicate the lowest average L_2-distance, while underlined values denote the second-best performing estimator in each scenario.

4.1 Background and dataset description

Our analysis utilises monthly total returns for WTI and Brent crude oil sourced from Bloomberg. Although the raw data begin in January 1980, data availability constraints

Table 2: **Summary of simulation study**

	Panel A ($m = 2$)		**Panel B** ($m = 10$)		**Panel C** ($m = 25$)		**Total (All Panels)**	
Model	**Best**	**2nd**	**Best**	**2nd**	**Best**	**2nd**	**Best**	**2nd**
OLS	0	0	9	0	8	5	**17**	**5**
RR	4	5	3	6	2	4	**9**	**15**
Liu	9	8	0	4	0	1	**9**	**13**
PR_c	2	2	3	9	9	5	**14**	**16**
PR_t	5	5	5	1	1	5	**11**	**11**
PR (Combined)	**7**	**7**	**8**	**10**	**10**	**10**	**25**	**27**
Total Scenarios	**20**	**20**	**20**	**20**	**20**	**20**	**60**	**60**

Notes. This table summarises the performance of all estimators across 60 independent scenarios, based on the results reported in Table 1. "Best" and "2nd" denote the number of times an estimator achieved the lowest and second-lowest average L_2-distance, respectively.

and the requirement that all factors be observed concurrently lead us to consider the following sample periods: (i) WTI, August 1988 to September 2024 (434 observations); and (ii) Brent, April 1998 to September 2024 (318 observations). A detailed technical discussion of benchmark selection and data source characteristics is provided in Appendix C.1.

Following the methodology of Sakkas and Tessaromatis (2020), we model monthly excess returns using a factor-based specification. The model is given by

$$y_{t+1} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{j,t} + \epsilon_{t+1}, \tag{32}$$

where y_{t+1} denotes the monthly excess return (over the risk-free rate) at time $t+1$, and $X_{j,t}$ represents the j-th normalised factor observed at time t.

For WTI, the model includes nine factors ($p = 9$): *Momentum*, *Basis*, *Basis Momentum*, *Skewness*, *Inflation Beta*, *Volatility*, *Hedging Pressure*, *Open Interest*, and *Value*. For Brent, *Hedging Pressure* is excluded due to data limitations, resulting in $p = 8$. To ensure comparability across variables measured in different units, all covariates are normalised to the range $[0, 1]$, while the response variable y remains unscaled. Definitions and economic motivations for each factor are provided in Appendix C.2, and the explicit construction formulas follow the framework established in Sakkas and Tessaromatis (2020).

Figures 2 and 3 present the monthly returns and cumulative performance of both bench-

marks. We apply the endogenous structural breakpoint test of Bai and Perron (2003) to segment the sample into distinct economic regimes. This procedure identifies 12 regimes for WTI and 11 for Brent, including major high-volatility episodes such as: (i) the Global Financial Crisis and subsequent recovery, spanning September 2008 to May 2011 for WTI and September 2008 to January 2011 for Brent; and (ii) the COVID-19 pandemic and its aftermath, covering March 2020 to July 2022 for WTI and February 2020 to July 2022 for Brent. A detailed timeline of these regimes and the associated market events is provided in the structural breakdown reported in Appendix C.3.

4.2 Data analysis

The OOS performance of all estimators is evaluated using an expanding training-window framework. This approach is anchored to the structural regimes identified in Section 4.1 and further detailed in Appendix C.3. For each transition from Period i to Period $i+1$, the five regression models are estimated using all available data up to the end of Period i. Their predictive performance is then evaluated over the testing window corresponding to the entirety of Period $i+1$.

Given a testing window of length n, observed excess returns y_t, and corresponding predictions $\hat{y}_t$, OOS performance is assessed using three standard metrics:

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2, \quad \text{RMSE} = \sqrt{\text{MSE}}, \quad \text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|. \tag{33}$$

MSE and Root Mean Squared Error (RMSE) impose a quadratic loss, thereby penalising large forecast errors more heavily and highlighting model instability during market shocks. In contrast, the Mean Absolute Error (MAE) applies a linear loss and is therefore more robust to extreme observations. While OLS and shrinkage estimators may experience variance inflation during turbulent market regimes, the PR framework incorporates a parity-based constraint designed to enhance predictive stability. A detailed discussion of the comparative stability of PR relative to penalised methods across these regimes is provided in Appendix C.4.

Tables 3 and 4 present the OOS predictive performance for WTI and Brent crude oil excess returns across 12 and 11 distinct economic periods, respectively. Although no single estimator consistently outperforms across all periods, the results demonstrate that our PR estimators provide greater stability during regimes of extreme market volatility and structural shifts, where traditional benchmarks often exhibit pronounced instability.

Table 3: **Performance comparison for WTI monthly returns**

Predicted	**MSE (%)**					**RMSE (%)**					**MAE (%)**				
Period	OLS	RR	Liu	PR_c	PR_t	OLS	RR	Liu	PR_c	PR_t	OLS	RR	Liu	PR_c	PR_t
2	63.83	6.68	10.65	**6.25**	7.02	79.89	25.85	32.63	**24.99**	26.49	72.22	**21.31**	26.33	22.08	24.16
3	**3.79**	3.90	3.87	3.81	3.83	**19.48**	19.76	19.67	19.53	19.57	**16.09**	16.62	16.41	16.42	16.47
4	13.98	**10.91**	12.07	11.47	12.20	37.40	**33.03**	34.74	33.86	34.92	29.14	**26.30**	27.49	26.74	27.52
5	**1.06**	2.20	1.76	1.20	2.22	**10.31**	14.82	13.26	10.95	14.89	**8.63**	12.38	11.23	9.23	12.45
6	**15.16**	86.42	83.18	63.44	81.76	**38.94**	92.96	91.21	79.65	90.42	**31.47**	82.77	81.57	69.84	80.21
7	7.86	1.45	2.29	2.02	**1.27**	28.04	12.04	15.13	14.20	**11.27**	23.96	9.52	13.10	11.86	**8.81**
8	9.44	1.50	1.41	**1.34**	1.46	30.73	12.24	11.88	**11.57**	12.10	26.95	9.42	9.24	**9.01**	9.32
9	191.20	0.86	18.00	**0.76**	172.71	138.28	9.26	42.43	**8.71**	131.42	52.93	7.75	22.09	**6.99**	51.41
10	10.11	2.75	**2.39**	2.76	4.08	31.80	16.58	**15.47**	16.62	20.21	23.77	13.33	**12.84**	13.36	15.48
11	45.64	9.39	11.33	**8.13**	8.41	67.56	30.65	33.66	**28.51**	29.00	57.45	20.92	23.95	20.40	**20.32**
12	26.42	**18.79**	19.03	18.79	18.80	51.40	**43.35**	43.62	43.35	43.35	47.98	**42.10**	42.43	42.10	42.10
Best Count	3	2	1	**4**	1	3	2	1	**4**	1	**3**	**3**	1	2	2
2nd Best	0	4	1	**5**	1	0	4	1	**5**	1	0	3	2	**6**	0

Notes. The table presents the MSE, RMSE, and MAE (in percentage points) defined in (33) for OLS, RR, Liu, PR_c and PR_t over 11 OOS predicted periods. For each period, the model with the best performance is highlighted in **bold red**, while the second-best is underlined. The bottom two rows summarise the total count of best and second-best performances achieved by each estimator across all periods.

For the WTI dataset, the summary rows indicate that PR_c attains the highest number of best and second-best MSE performances across the evaluated periods. Examining specific high-volatility events, the performance of the estimators varies. During the Global Financial Crisis (Period 6), all methods struggle, with OLS achieving the lowest MSE of 15.16%; however, PR_c secures the second-best overall position, notably outperforming the other shrinkage estimators. This illustrates that even when standard shrinkage methods falter, the parity constraint enhances predictive robustness. In the subsequent recovery phase (Period 7), PR_t attains the lowest MSE of 1.27%, outperforming all other competitors. During the COVID-19 pandemic (Period 11), the parity-based framework demonstrates structural resilience: PR_c achieves the minimum MSE of 8.13%, while PR_t ranks second-best at 8.41%, whereas RR,

Table 4: **Performance comparison for Brent monthly returns**

Predicted	**MSE (%)**					**RMSE (%)**					**MAE (%)**				
Period	OLS	RR	Liu	PR_c	PR_t	OLS	RR	Liu	PR_c	PR_t	OLS	RR	Liu	PR_c	PR_t
2	13.06	9.08	**8.36**	9.27	9.02	36.14	30.13	**28.91**	30.45	30.03	31.92	26.69	**25.47**	27.01	26.60
3	**1.90**	2.03	1.96	2.01	2.02	**13.79**	14.24	13.99	14.17	14.20	**11.16**	11.99	11.72	11.81	11.88
4	**0.74**	1.13	0.77	1.02	1.21	**8.62**	10.63	8.80	10.11	11.02	**6.89**	8.69	6.90	8.09	8.68
5	70.09	153.14	149.59	**43.99**	191.14	83.72	123.75	122.31	**66.33**	138.25	80.43	111.96	110.77	**63.22**	122.86
6	36.90	**5.28**	13.16	6.83	8.32	60.75	**22.98**	36.27	26.13	28.84	29.01	**13.74**	18.64	14.80	15.67
7	1.37	1.56	1.37	**1.19**	1.22	11.69	12.48	11.69	**10.92**	11.05	8.99	10.72	9.76	**8.90**	9.10
8	16.42	5.53	**2.77**	5.12	4.42	40.52	23.52	**16.66**	22.62	21.02	33.75	19.18	**13.68**	18.44	17.18
9	4.47	2.47	2.50	**2.44**	2.64	21.14	15.72	15.80	**15.62**	16.24	18.09	13.82	13.87	**13.70**	14.41
10	265.82	37.83	53.16	**8.09**	163.11	163.04	61.51	72.91	**28.44**	127.71	110.94	40.70	49.40	**25.62**	81.05
11	47.19	**18.23**	23.42	18.25	18.25	68.70	**42.69**	48.40	42.71	42.71	67.32	**41.45**	47.32	41.48	41.48
Best Count	2	2	2	**4**	0	2	2	2	**4**	0	2	2	2	**4**	0
2nd Best	1	2	2	2	**4**	1	2	2	2	**4**	2	2	2	2	**3**

Notes. The table presents the MSE, RMSE, and MAE (in percentage points) defined in (33) for OLS, RR, Liu, PR_c and PR_t over 10 OOS predicted periods. For each period, the model with the best performance is highlighted in **bold red**, while the second-best is underlined. The bottom two rows summarise the total count of best and second-best performances achieved by each estimator across all periods.

Liu, and OLS fail to maintain comparable stability.

A similar pattern is observed for the Brent dataset, where the summary counts indicate that PR_c again secures the highest number of best MSE performances, while PR_t achieves the most second-best rankings. During the Financial Crisis (Period 5), PR_c emerges as the most robust estimator, delivering the lowest MSE of 43.99%. In contrast, traditional shrinkage methods such as RR and Liu suffer from substantial errors, performing even worse than the second-best OLS. During the subsequent recovery phase (Period 6), RR attains the lowest MSE at 5.28%, with PR_c closely following as the second-best model at 6.83%. Most notably, during the extreme volatility of the COVID-19 pandemic (Period 10), PR_c proves highly reliable, achieving a remarkably low MSE of 8.09%. In this extreme regime, conventional methods fail: OLS produces an MSE of 265.82%, while RR and Liu incur errors of 37.83% and 53.16%, respectively.

Consistent with our findings from the simulation study, the relative performance of PR_t and PR_c varies across market regimes. While PR_t exhibits particular strength in capturing trends during the WTI recovery phase, PR_c demonstrates superior robustness in the most volatile and ill-conditioned periods, such as the Financial Crisis and the COVID-19 pandemic,

across both commodities. By employing a risk-parity structure, these estimators maintain stability even when factor correlations deteriorate abruptly. Overall, the results presented in Tables 3 and 4 underscore that parity-based regression provides a critical mechanism for ensuring parameter stability, rendering the PR framework a more reliable alternative to conventional shrinkage methods, such as RR or Liu, when forecasting oil returns under severe global shocks.

5 Conclusions

We have introduced Parity Regression, a novel multiple linear regression framework that replaces aggregate error minimisation with an elasticity-based principle that distributes prediction error evenly across model parameters. This formulation induces a structurally balanced regularisation mechanism that is particularly well suited to noisy environments, including time series settings characterised by structural shifts and evolving dynamics. We provided a rigorous theoretical characterisation of the new estimator, establishing its existence, uniqueness, and structural properties, and demonstrated that it can be embedded within a unified Generalised Weighted Mean Constrained framework that also encompasses classical penalised and shrinkage estimators.

By reinterpreting regularisation through elasticity balancing rather than norm penalisation alone, Parity Regression offers a conceptually distinct yet mathematically coherent extension of existing methodology. Theoretical guarantees are corroborated by simulation studies and real-data applications, which confirm its stability and competitive performance. Collectively, these results position Parity Regression as a substantive methodological advancement with strong foundations for further analytical and high-dimensional development.

Acknowledgements

The authors express their sincere gratitude to Professor Alexandru Bădescu (University of Calgary) for his valuable comments and constructive guidance throughout the theoretical development and implementation phases of this research. His insights have materially contributed to the rigour and clarity of the final manuscript.

References

Asimit, A. V., Furman, E., Tang, Q., and Vernic, R. (2011). Asymptotics for risk capital allocations based on conditional tail expectation. *Insurance: Mathematics and Economics*, 49(3):310–324.

Asimit, A. V., Vernic, R., and Zitikis, R. (2013). Evaluating risk measures and capital allocations based on multi-losses driven by a heavy-tailed background risk: The multivariate pareto-ii model. *Risks*, 1(1):14–33.

Asimit, V., Chen, Z., and Lassance, N. (2026a). Distribution-free shrinkage of high-dimensional mean vector. *Journal of Business & Economic Statistics, Forthcoming*.

Asimit, V., Chong, W. F., Tunaru, R., and Zhou, F. (2025). Portfolio selection and risk sharing via risk budgeting. *Insurance: Mathematics and Economics*, page 103139.

Asimit, V., Cidota, M. A., Chen, Z., and Asimit, J. (2026b). Slab and shrinkage linear regression estimation. Preprint.

Asimit, V., Peng, L., Tunaru, R., and Zhou, F. (2026c). Risk budgeting under general risk measures. Preprint.

Asimit, V., Peng, L., Wang, R., and Yu, A. (2019). An efficient approach to quantile capital allocation and sensitivity analysis. *Mathematical Finance*, 29(4):1131–1156.

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, 68(3):929–985.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Bakshi, G., Gao, X., and Rossi, A. G. (2019). Understanding the sources of risk underlying the cross section of commodity returns. *Management Science*, 65:459–954.

Bessembinder, H. (1992). Systematic risk, hedging pressure, and risk premiums in futures markets. *The Review of Financial Studies*, 5(4):637–667.

Bodnar, O., Bodnar, T., and Parolya, N. (2022). Recent advances in shrinkage-based high-dimensional inference. *Journal of Multivariate Analysis*, 188.

Boons, M. and Prado, M. S. (2019). Basis momentum. *The Journal of Finance*, 74(1):239–279.

Chen, S. and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1.

Fernandez-Perez, A., Frijns, B., Fuertes, A.-M., and Miffre, J. (2018). The skewness of commodity futures returns. *Journal of Banking & Finance*, 86:143–158.

Gauss, C. F. (1821). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Henricus Dieterich, Göttingen.

Gorton, G. and Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2):47–68.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hong, H. and Yogo, M. (2012). What does futures market interest tell us about the macroeconomy and asset prices? *Journal of Financial Economics*, 105:473–490.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1:361–379.

Keynes, J. M. (1930). *A Treatise on Money*. Macmillan.

Liu, K. (1993). A new class of blased estimate in linear regression. *Communications in Statistics-Theory and Methods*, 22(2):393–402.

Liu, K. (2003). Using liu-type estimator to combat collinearity. *Communications in Statistics - Theory and Methods*, 32(5):1009–1020.

Markov, A. A. (1912). *Wahrscheinlichkeitsrechnung*. B. G. Teubner, Leipzig.

Miffre, J. and Rallis, G. (2007). Momentum strategies in commodity futures markets. *Journal of Banking & Finance*, 31(6):1863–1886.

Sakkas, A. and Tessaromatis, N. (2020). Factor based commodity investing. *Journal of Banking & Finance*, 115:105782.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1:197–206.

Tasche, D. (1999). Risk contributions and performance measurement. *Report of the Lehrstuhl für mathematische Statistik, TU München*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tikhonov, A. N. et al. (1943). On the stability of inverse problems. In *Dokl. akad. nauk sssr*, volume 39, pages 195–198.

Working, H. (1949). The theory of price of storage. *The American Economic Review*, 39(6):1254–1262.

Yang, F. (2013). Investment shocks and the commodity basis spread. *Journal of Financial Economics*, 110:164–184.

A Proofs

A.1 Proof of Theorem 1

We first prove part i). The first-order necessary conditions for any solution $\boldsymbol{\theta}^\star$ of (12) leads to

$$\theta_k^\star \frac{\partial \mathcal{L}}{\partial \theta_k}(\boldsymbol{\theta}^\star) = \mu, \quad \text{for all } 1 \le k \le q,$$

which further implies that $\boldsymbol{\theta}^\star$ satisfies (11).

For part ii), we first demonstrate that (12) has a solution. Let $f(\boldsymbol{\theta}; \mu)$ be the objective function of (12). We begin by showing that

$$\lim_{\|\boldsymbol{\theta}\|_\infty \to \infty} f(\boldsymbol{\theta}; \mu) = +\infty \quad \text{for any } \mu > 0. \tag{34}$$

Indeed,

$$\begin{aligned} f(\boldsymbol{\theta}; \mu) &= \mathcal{L}(\boldsymbol{\theta}) - \mu \sum_{k=1}^{q} \log(\delta_k \theta_k) \\ &\ge M_1 \|\boldsymbol{\theta}\|_\infty - \mu \sum_{k=1}^{q} \log\left(\frac{\delta_k \theta_k}{\|\boldsymbol{\theta}\|_\infty} \right) - \mu q \log \|\boldsymbol{\theta}\|_\infty \quad \text{for } \|\boldsymbol{\theta}\|_\infty \text{ is sufficiently large} \end{aligned}$$

$$\geq M_1\|\boldsymbol{\theta}\|_\infty - \mu q \log \|\boldsymbol{\theta}\|_\infty$$
$$\to \infty \quad \text{as} \quad \|\boldsymbol{\theta}\|_\infty \to \infty.$$

Next, we show that

$$\lim_{\|\boldsymbol{\theta}\|_{-\infty}\to 0} f(\boldsymbol{\theta};\mu) = +\infty, \qquad \text{for any } \mu > 0,$$

where $\|\boldsymbol{\theta}\|_{-\infty} = \min_i |\theta_i|$. This follows from

$$\lim_{\|\boldsymbol{\theta}\|_{-\infty}\to 0} \sum_{k=1}^{q} \log(\delta_k\theta_k) = -\infty$$

and the fact that $\mathcal{L}(\boldsymbol{\theta}) > 0$ for any $\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta})$. Thus, there exists $\epsilon > 0$ and $K > 0$ such that

$$\inf_{\boldsymbol{\theta}\in\mathcal{K}_q(\boldsymbol{\delta})} f(\boldsymbol{\theta};\mu) = \inf_{\boldsymbol{\theta}\in\mathcal{B}_{\epsilon,K}} f(\boldsymbol{\theta};\mu),$$

where

$$\mathcal{B}_{\epsilon,K} = \{\boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta}) : \|\boldsymbol{\theta}\|_{-\infty} \geq \epsilon, \|\boldsymbol{\theta}\|_\infty \leq K\}.$$

The conclusion follows since $\mathcal{B}_{\epsilon,K}$ is compact and f is continuous.

Finally, note that $f(\boldsymbol{\theta};\mu)$ of is strictly convex in $\boldsymbol{\theta}$ since $\mu > 0$, $\mathcal{L}$ is convex and $\log(\delta_k\theta_k)$ is strictly concave in θ_k for all k. Hence, (12) has a unique solution.

For part iii), we begin by proving that (12) admits a unique solution. Since the proof is the same as part ii) except for (34), we only show (34) holds. Since $\mathcal{L}$ is positive and homogeneous of order τ, then for any $0 < \zeta < \tau$, we have that $\mathcal{L}(\boldsymbol{\theta}) \geq \|\boldsymbol{\theta}\|_\infty^{\tau-\zeta}$ for those $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}\|_\infty$ is sufficiently large. Thus,

$$f(\boldsymbol{\theta};\mu) = \mathcal{L}(\boldsymbol{\theta}) - \mu \sum_{k=1}^{q} \log(\delta_k\theta_k) \tag{35}$$
$$\geq C\|\boldsymbol{\theta}\|_\infty^\tau - \mu q \log \|\boldsymbol{\theta}\|_\infty \quad \text{for } \|\boldsymbol{\theta}\|_\infty \text{ is sufficiently large}$$
$$\to \infty \quad \text{as} \quad \|\boldsymbol{\theta}\|_\infty \to \infty,$$

which gives the needed result.

We now show that $\mathcal{L}(\boldsymbol{\theta}^\star(\mu)) = \frac{q\mu}{\tau}$ for all $\mu > 0$. Since $\boldsymbol{\theta}^\star(\mu)$ is a solution of (12), the first-order conditions give

$$\theta_k \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k} = \mu \,\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star(\mu)} \quad \text{for any } 1 \le k \le q$$

The latter and Euler's Homogeneous Function Theorem yield $\mathcal{L}(\boldsymbol{\theta}^\star(\mu)) = q\mu/\tau$.

Next, we show that $\boldsymbol{\theta}^\star(\mu) = \mu^{1/\tau}\boldsymbol{\theta}^\star(1)$ for all $\mu > 0$. The homogeneity of $\mathcal{L}$ implies that

$$f\big(\mu^{-1/\tau}\boldsymbol{\theta}; 1\big) = \frac{1}{\mu} f(\boldsymbol{\theta}; \mu) + \frac{q}{\tau}\log\mu \quad \text{for any } \boldsymbol{\theta} \in \mathcal{K}_q(\boldsymbol{\delta}) \text{ and } \mu > 0,$$

which further leads to

$$\operatorname*{argmin}_{\boldsymbol{\theta}} f\big(\mu^{-1/\tau}\boldsymbol{\theta}; 1\big) = \operatorname*{argmin}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \mu) = \boldsymbol{\theta}^\star(\mu),$$

as $\mu > 0$. It can be easily seen that

$$\mu^{\frac{1}{\tau}}\theta^\star(1) = \operatorname*{argmin}_{\boldsymbol{\theta}} f\big(\mu^{-1/\tau}\boldsymbol{\theta}; 1\big)$$

which shows that $\boldsymbol{\theta}^\star(\mu) = \mu^{1/\tau}\boldsymbol{\theta}^\star(1)$.

Lastly, we prove that the final statement of part iii). if $\widetilde{\boldsymbol{\theta}} = \mu_0^{1/\tau}\boldsymbol{\theta}^\star(1)$ for some $\mu_0 > 0$ then $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star(\mu_0)$ and by part i), $\widetilde{\boldsymbol{\theta}}$ is a parity estimator. Conversely, assume that $\widetilde{\boldsymbol{\theta}}$ is a parity estimator so that there exists $\mu_0 > 0$ such that

$$\theta_k \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_k} \,\big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} = \mu_0 \quad \text{for all } 1 \le k \le q. \tag{36}$$

Euler's Theorem gives $\mathcal{L}(\widetilde{\boldsymbol{\theta}}) = q\mu_0/\tau = \mathcal{L}(\boldsymbol{\theta}(\mu_0))$. Thus $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mu_0) = \mu_0^{1/\tau}\boldsymbol{\theta}^\star(1)$ by the uniqueness of the solution of (12). The proof is now complete.

A.2 Proof of Theorem 2

We begin by proving part i). Based on the proof of Theorem 1, the optimisation problem in (20) has a unique solution $\big((\boldsymbol{\theta}^\star(\lambda, t, \mu))^\top, \theta^\star_{p+1}(\lambda, t, \mu)\big)^\top$ for any parameters (λ, t, μ) satisfying $\lambda \ge 0$, $t \ge 0$, and $\mu > 0$. This solution also fulfills the parity conditions specified in (11).

Define $\mathcal{RRSSC}_k$ for each k as follows

$$\mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)=\theta_k\frac{\partial\mathcal{RRSS}\big(\boldsymbol{\theta}(\lambda,t,\mu),\theta_{p+1}(\lambda,t,\mu);\lambda\big)}{\partial\theta_k}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star(\lambda,t,\mu)},$$

for all $0\le k\le p+1$. Since $\big((\boldsymbol{\theta}^\star(\lambda,t,\mu))^\top,\theta^\star_{p+1}(\lambda,t,\mu)\big)^\top$ is a unique solution, it must satisfy the stationary conditions

$$\begin{cases}\mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)=\mu, & \text{for } k=0,\dots,p,\\ \mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)=\mu t, & \text{for } k=p+1.\end{cases}\tag{37}$$

Applying Euler's homogeneous function theorem, and $\mathcal{RRSS}$ being a homogeneous function of order $\tau=2$, we can express

$$\begin{aligned}\mathcal{RRSS}\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)&=\frac{1}{2}\sum_{k=0}^{p+1}\mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)\\&=\frac{1}{2}\left(\sum_{k=0}^{p}\mu+\mu t\right)=\frac{(p+1+t)\mu}{2}.\end{aligned}$$

We find that for each $0\le k\le p$,

$$\mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big)=\left(\frac{2}{p+1+t}\right)\mathcal{RRSS}\Big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu);\lambda\Big).$$

Since $\mathcal{RRSS}(\boldsymbol{\theta},\theta_{p+1};\lambda)$ is homogeneous, $\mathcal{RRSSC}_k$ is also homogeneous of the same order, and in turn, the following holds for all $0\le k\le p$ and any $m>0$

$$\mathcal{RRSSC}_k\Big(m\big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu)\big);\lambda\Big)=\frac{2\mathcal{RRSS}\Big(m\big(\boldsymbol{\theta}^\star(\lambda,t,\mu),\theta^\star_{p+1}(\lambda,t,\mu)\big);\lambda\Big)}{p+1+t}.\tag{38}$$

Setting $m=1/\theta^\star_{p+1}(\lambda,t,\mu)$, we conclude that $\big((\boldsymbol{\theta}^\star(\lambda,t,\mu))^\top,\theta^\star_{p+1}(\lambda,t,\mu)\big)^\top/\theta^\star_{p+1}(\lambda,t,\mu)$ is the unique PR estimate as defined in (21) within $\mathcal{K}_{p+2}(\boldsymbol{\delta})$, thereby satisfying the parity estimator in (13).

To show that $\big((\boldsymbol{\theta}^\star(\lambda,t,\mu))^\top,\theta^\star_{p+1}(\lambda,t,\mu)\big)^\top/\theta^\star_{p+1}(\lambda,t,\mu)$ is constant with respect to $\mu>0$ for any given (λ,t), consider the objective function of (20), denoted as $H(\boldsymbol{\theta},\theta_{p+1};\lambda,t,\mu)$. We

find that

$$H\Big(\mu^{-1/2}\boldsymbol{\theta}, \mu^{-1/2}\theta_{p+1}; \lambda, t, 1\Big) = \frac{1}{\mu} H\big(\boldsymbol{\theta}, \theta_{p+1}; \lambda, t, \mu\big) - \left(\frac{p+1+t}{2}\right) \log \mu,$$

which implies

$$\underset{(\boldsymbol{\theta}, \theta_{p+1})}{\operatorname{argmin}} H\Big(\mu^{-1/2}\boldsymbol{\theta}, \mu^{-1/2}\theta_{p+1}; \lambda, t, 1\Big) = \underset{(\boldsymbol{\theta}, \theta_{p+1})}{\operatorname{argmin}} H\big(\boldsymbol{\theta}, \theta_{p+1}; \lambda, t, \mu\big), \tag{39}$$

yielding a unique solution $\big(\boldsymbol{\theta}^\star(\lambda, t, \mu)^\top, \theta^\star_{p+1}(\lambda, t, \mu)\big)^\top$ for $\mu > 0$ and $t \geq 0$. Since

$$\underset{(\boldsymbol{y}, y_{p+1})}{\operatorname{argmin}} f\big(\boldsymbol{y}, y_{p+1}; \lambda, t, 1\big) = \big(\boldsymbol{\theta}^\star(\lambda, t, 1)^\top, \theta^\star_{p+1}(\lambda, t, 1)\big)^\top,$$

where $\big(\boldsymbol{\theta}^\star(\lambda, t, 1)^\top, \theta^\star_{p+1}(\lambda, t, 1)\big)^\top$ is the optimal solution in (20) with $\tilde{\mu} = 1$. Together with (39), we obtain

$$\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, \mu)\big)^\top, \theta^\star_{p+1}(\lambda, t, \mu)\Big)^\top = \mu^{1/2}\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, 1)\big)^\top, \theta^\star_{p+1}(\lambda, t, 1)\Big)^\top \quad \text{for any} \quad \mu > 0, \tag{40}$$

Thus,

$$\frac{\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, \mu)\big)^\top, \theta^\star_{p+1}(\lambda, t, \mu)\Big)^\top}{\theta^\star_{p+1}(\lambda, t, \mu)} = \frac{\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, 1)\big)^\top, \theta^\star_{p+1}(\lambda, t, 1)\Big)^\top}{\theta^\star_{p+1}(\lambda, t, 1)} \quad \text{for any} \quad \mu > 0, \tag{41}$$

showing that our PR estimates do not depend on the normalising constant μ. Choosing $\mu^\star > 0$ such that $\theta^\star_{p+1}(\lambda, t, \mu^\star) = 1$, then (40) leads to the required result displayed in (23). Furthermore, (40) and (41) yield

$$\big(\boldsymbol{\theta}^\star(\lambda, t, \mu^\star)^\top, 1\big)^\top = (\mu^\star)^{1/2}\Big(\big(\boldsymbol{\theta}^\star(\lambda, t, 1)\big)^\top, \theta^\star_{p+1}(\lambda, t, 1)\Big)^\top = \frac{\big((\boldsymbol{\theta}^\star(\lambda, t, 1))^\top, \theta^\star_{p+1}(\lambda, t, 1)\big)^\top}{\theta^\star_{p+1}(\lambda, t, 1)},$$

which concludes that $\mu^\star = \big(\theta^\star_{p+1}(\lambda, t, 1)\big)^{-2}$. The rest of the proof of part i) is straightforward, since it relies on arguments similar to those above.

Next, we show the proof of part ii). We begin by noting that (24) is a convex optimisation problem because the objective function $\mathcal{RRSS}(\boldsymbol{\theta}, \theta_{p+1}; \lambda)$ is strictly convex in $(\boldsymbol{\theta}, \theta_{p+1})$. To establish the existence of a solution, we observe that: i) the objective function grows

unboundedly near infinity (as shown in the proof of Theorem 1), and ii) any point $(\boldsymbol{\theta}, \theta_{p+1}) \in \mathcal{K}_{p+2}(\boldsymbol{\delta})$ becomes infeasible when $\|\theta_k\|_{-\infty} \leq \epsilon$ for sufficiently small ϵ, since $\lim_{t \to 0^+} \log t = -\infty$. This implies that the feasible set $\mathcal{K}_{p+2}(\boldsymbol{\delta})$ excludes boundary points where any θ_k approaches zero, effectively restricting the solution to a bounded subset of $\mathcal{K}_{p+2}(\boldsymbol{\delta})$. Given this compactness and the strict convexity of $\mathcal{RRSS}(\boldsymbol{\theta}, \theta_{p+1}; \lambda)$, a solution is guaranteed to exist.

We now proceed to show that (24) admits a unique solution, which requires two main steps.

First, assume that $\big(\mathbf{z}^{*T}, z^{\star}_{p+1}\big)$ is an optimal solution of (24) for which the inequality constraint in (24) is binding (becomes an identity). If the constraint would not have been binding, we define

$$\kappa^{\star} = \exp\left\{ \tilde{\mu} - \sum_{k=0}^{p} \log(\delta_k z^{\star}_k) - t \log\big(\delta_{p+1} z^{\star}_{p+1}\big) \right\},$$

and note that $0 < \kappa^{\star} < 1$, as the constraint would be non-binding at $\big(\mathbf{z}^{*\top}, z^{\star}_{p+1}\big)$. Then,

$$\mathcal{RRSS}\big(\kappa^{\star}\mathbf{z}^{\star}, \kappa^{\star} z^{\star}_{p+1}; \lambda\big) = \big(\kappa^{\star}\big)^2 \mathcal{RRSS}\big(\mathbf{z}^{\star}, z^{\star}_{p+1}; \lambda\big) < \mathcal{RRSS}\big(\mathbf{z}^{\star}, z^{\star}_{p+1}; \lambda\big),$$

where the first equality is due to the homogeneity of order 2 of $\mathcal{RRSS}(\cdot; \lambda)$ on $\mathcal{K}_{p+2}(\boldsymbol{\delta})$, and the inequality follows from $0 < \kappa^{\star} < 1$, (17), and (18). This contradicts our assumption that $\big(\mathbf{z}^{*T}, z^{\star}_{p+1}\big)$ is a solution of (24), implying that any solution of (24) must satisfy the constraint as an identity.

Next, assume that there exist two distinct solutions, $\big(\mathbf{z}^{*T}, z^{\star}_{p+1}\big)$ and $\big(\mathbf{z}^{**T}, z^{\star\star}_{p+1}\big)$, for a given tuple $(\lambda, t, \tilde{\mu})$. Define

$$\big(\mathbf{z}^{\star\star\star}, z^{\star\star\star}_{p+1}\big) = \gamma\big(\mathbf{z}^{*T}, z^{\star}_{p+1}\big) + (1-\gamma)\big(\mathbf{z}^{**T}, z^{\star\star}_{p+1}\big) \quad \text{where } 0 < \gamma < 1.$$

Since (24) is a convex problem, $\big(\mathbf{z}^{\star\star\star}, z^{\star\star\star}_{p+1}\big)$ solves (24). Now, we find

$$\sum_{k=0}^{p} \log(\delta_k z^{\star\star\star}_k) + t \log\big(\delta_{p+1} z^{\star\star\star}_{p+1}\big) \tag{42}$$

$$
\begin{aligned}
&> \gamma\left(\sum_{k=0}^{p}\log(\delta_k z_k^\star)+t\log\big(\delta_{p+1}z_{p+1}^\star\big)\right)+(1-\gamma)\left(\sum_{k=0}^{p}\log(\delta_k z_k^{\star\star})+t\log\big(\delta_{p+1}z_{p+1}^{\star\star}\big)\right)\\
&= \tilde{\mu},
\end{aligned}
$$

since $\log(\cdot)$ is strictly concave on $\mathbb{R}_+^\star$ and both $\big(\mathbf{z}^{*T}, z_{p+1}^\star\big)$ and $\big(\mathbf{z}^{**T}, z_{p+1}^{\star\star}\big)$ satisfy the constraint as an identity. However, since $\big(\mathbf{z}^{\star\star\star}, z_{p+1}^{\star\star\star}\big)$ solves (24), it must also satisfy the constraint as an identity, leading to a contradiction with (42). This completes the proof of uniqueness for (24).

Further, by Euler's homogeneous function theorem and the Karush-Kuhn-Tucker (KKT) conditions, we have

$$
\mathcal{RRSSC}_k\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}),\theta_{p+1}^{\star\star}(\lambda,t,\tilde{\mu});\lambda\big)=\left(\frac{2}{p+1+t}\right)\mathcal{RRSS}\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}),\theta_{p+1}^{\star\star}(\lambda,t,\tilde{\mu});\lambda\big),
$$

which, together with the equivalent variant of (38), implies that

$$
\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top,\theta_{p+1}^{\star\star}(\lambda,t,\tilde{\mu})\Big)^\top/\theta_{p+1}^{\star\star}\big(\lambda,t,\tilde{\mu}\big)
$$

is the unique PR estimate as defined in (25) within $\mathcal{K}_{p+2}(\boldsymbol{\delta})$, fulfilling the condition in (13).

We now demonstrate that $\big((\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}))^\top,\theta_{p+1}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top/\theta_{p+1}^{\star\star}\big(\lambda,t,\tilde{\mu}\big)$ is constant with respect to $\tilde{\mu}\in\mathbb{R}$ for any fixed (λ,t). For any tuple $(\lambda,t,\tilde{\mu})$, the unique solution of (24), denoted by $\big((\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}))^\top,\theta_{p+1}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top$, is equivalent to solving

$$
\begin{cases}
\min\limits_{(\boldsymbol{\theta},\theta_{p+1})\in\mathcal{K}_{p+2}(\boldsymbol{\delta})} & \mathcal{RRSS}\Big(e^{\frac{\tilde{\mu}}{p+1+t}}\boldsymbol{\theta},e^{\frac{\tilde{\mu}}{p+1+t}}\theta_{p+1};\lambda\Big)\\
\text{s.t.} & \sum\limits_{k=0}^{p}\log\Big(\delta_k e^{\frac{\tilde{\mu}}{p+1+t}}\theta_k\Big)+t\log\Big(\delta_{p+1}e^{\frac{\tilde{\mu}}{p+1+t}}\theta_{p+1}\Big)\geq 0
\end{cases}
\tag{43}
$$

This problem admits a unique solution

$$
\Big(e^{\frac{\tilde{\mu}^\star}{p+1+t}}\boldsymbol{\theta}^{\star\star}(\lambda,t,0)^\top,e^{\frac{\tilde{\mu}^\star}{p+1+t}}\theta_{p+1}^{\star\star}(\lambda,t,0)\Big)^\top
$$

for $\tilde{\mu}=0$ and $t\geq 0$. The uniqueness of the solution for (24) and (43) implies that

$$
\Big(\big(\boldsymbol{\theta}^{\star\star}\big(\lambda,t,\tilde{\mu}\big)\big)^\top,\theta_{p+1}^{\star\star}\big(\lambda,t,\tilde{\mu}\big)\Big)^\top=e^{\frac{\tilde{\mu}^\star}{p+1+t}}\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,0)\big)^\top,\theta_{p+1}^{\star\star}(\lambda,t,0)\Big)^\top\quad\text{for all }\tilde{\mu}\in\mathbb{R}.\tag{44}
$$

Thus,

$$\frac{\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu})\big)^\top,\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})\Big)^\top}{\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})}=\frac{\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,0)\big)^\top,\theta^{\star\star}_{p+1}(\lambda,t,0)\Big)^\top}{\theta^{\star\star}_{p+1}(\lambda,t,0)}\quad\text{for all }\tilde{\mu}\in\mathbb{R}. \tag{45}$$

This confirms that PR estimates are independent of the normalising constant $\tilde{\mu}$. By selecting $\tilde{\mu}^\star$ so that $\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu}^\star)=1$, we arrive at the desired result in (26), which is a straightforward consequence of (44). Moreover, (44) and (45) give that

$$\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}^\star)^\top,0\big)^\top=e^{\frac{\tilde{\mu}^\star}{p+1+t}}\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,0)\big)^\top,\theta^{\star\star}_{p+1}(\lambda,t,0)\Big)^\top=\frac{\Big(\big(\boldsymbol{\theta}^{\star\star}(\lambda,t,0)\big)^\top,\theta^{\star\star}_{p+1}(\lambda,t,0)\Big)^\top}{\theta^{\star\star}_{p+1}(\lambda,t,0)},$$

which concludes that $\tilde{\mu}^\star=-(p+1+t)\log\big(\theta^{\star\star}_{p+1}(\lambda,t,0)\big)$. Further, we note that the Slater's condition is satisfied in (24), and therefore, the strong duality in (24) holds.

Finally, we proceed with the proof of part iii). Using the notation introduced in parts i) and ii), recall the PR estimates in (21) and (25) satisfy

$$\widehat{\boldsymbol{\beta}}^{PR}(\lambda,t)=\frac{\boldsymbol{\theta}^\star(\lambda,t,\mu)}{\theta^\star_{p+1}(\lambda,t,\mu)}\quad\text{and}\quad\widehat{\widehat{\boldsymbol{\beta}}}^{PR}(\lambda,t)=\frac{\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu})}{\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})} \tag{46}$$

for any $\mu>0$ and $\tilde{\mu}\in\mathbb{R}$. Since strong duality holds in (24), let $\gamma^\star$ be the dual optimal multiplier in (24) associated with the logarithmic constraint

$$\sum_{k=0}^{p}\log(\delta_k\theta_k)+t\log(\theta_{p+1})\geq\tilde{\mu}.$$

Then, (24) is equivalent to minimising the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta},\theta_{p+1};\lambda,t,\tilde{\mu};\gamma)=\mathcal{RRSS}\big(\boldsymbol{\theta},\theta_{p+1};\lambda\big)-\gamma\left(\sum_{k=0}^{p}\log(\delta_k\theta_k)+t\log(\delta_{p+1}\theta_{p+1})-\tilde{\mu}\right), \tag{47}$$

and the KKT conditions imply that $\gamma=\gamma^\star$, ensuring that the constraint is active, i.e.,

$$\sum_{k=0}^{p}\log(\delta_k\theta^{\star\star}_k)+t\log\big(\delta_{p+1}\theta^{\star\star}_{p+1}\big)=\tilde{\mu}.$$

Furthermore, the stationarity conditions for (47) yield

$$\begin{cases} \mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}),\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu});\lambda\Big)=\gamma^\star, & \text{for } k=0,\ldots,p,\\ \mathcal{RRSSC}_k\Big(\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}),\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu});\lambda\Big)=\gamma^\star t, & \text{for } k=p+1. \end{cases} \tag{48}$$

Therefore, (37) and (48) imply that solving the primal in (24) is equivalent to solving (20) with $\mu=\gamma^\star$. There exists $\alpha^\star>0$ such that

$$\big((\boldsymbol{\theta}^\star(\lambda,t,\mu))^\top,\theta^\star_{p+1}(\lambda,t,\mu)\big)^\top=\alpha^\star\big((\boldsymbol{\theta}^{\star\star}(\lambda,t,\tilde{\mu}))^\top,\theta^{\star\star}_{p+1}(\lambda,t,\tilde{\mu})\big)^\top \tag{49}$$

for any $\mu>0$ and $\tilde{\mu}\in\mathbb{R}$. Equations (46) and (49) conclude that $\hat{\boldsymbol{\beta}}^{PR}(\lambda,t)=\widehat{\widehat{\boldsymbol{\beta}}}^{PR}(\lambda,t)$. The proof is now complete.

A.3 Proof of Proposition 2

The proofs of parts i) and ii) are very similar, and thus, we only show part i). Theorem 2 i) tells us that there exists $\mu^\star>0$ such that $\big((\widehat{\boldsymbol{\beta}}^{PR}(\lambda,t))^\top,1\big)^\top$ uniquely solves (20) with $\mu=\mu^\star$, and in turn we have that

$$\begin{aligned}&\mathcal{RRSS}\Big(\widehat{\boldsymbol{\beta}}^{PR}(\lambda,t),1;\lambda\Big)-\mu^\star\sum_{k=0}^{p}\log\Big(\delta_k\widehat{\boldsymbol{\beta}}^{PR}_k(\lambda,t)\Big)\\ &<\mathcal{RRSS}\Big(\widehat{\boldsymbol{\beta}}^{RR}(\lambda),1;\lambda\Big)-\mu^\star\sum_{k=0}^{p}\log\Big(\delta_k\widehat{\boldsymbol{\beta}}^{RR}_k(\lambda)\Big).\end{aligned}$$

Consequently,

$$0\le\mathcal{RRSS}\Big(\widehat{\boldsymbol{\beta}}^{PR}(\lambda,c),1;\lambda\Big)-\mathcal{RRSS}\Big(\widehat{\boldsymbol{\beta}}^{RR}(\lambda),1;\lambda\Big)<\mu^\star\sum_{k=0}^{p}\log\left(\frac{\widehat{\boldsymbol{\beta}}^{PR}_k(\lambda,t)}{\widehat{\boldsymbol{\beta}}^{RR}_k(\lambda)}\right),$$

where the first inequality is due to (4), and in turn we could conclude (27) as $\mu^\star,t\ge0$. The proof of part i) is now complete, which completes the entire proof.

B Data generation process for synthetic data

This section provides a detailed description of the DGP employed in the numerical experiments presented in Section 3. The design aims to assess the performance of the various estimators under controlled levels of multicollinearity and heteroscedasticity. For reproducibility, the procedure is structured into three main steps, which are outlined below.

Step 1: Feature Generation

i) **Correlation Matrix Construction**: We construct a symmetric correlation matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$ for m features. The elements are defined as $\Sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$ and $\Sigma_{ii} = 1$ otherwise, simulating multicollinearity. The correlation parameter ρ is varied across $\{-0.75, -0.5, 0, 0.5, 0.75\}$ to assess the models under different levels of association.

ii) **Feature Vector Simulation**: Feature vectors $\mathbf{X}_i$ are generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Each feature is subsequently standardised to ensure zero mean and unit variance for each individual feature.

Step 2: Response Variable Generation

i) **Regression Parameters**: We generate the true regression parameters $\boldsymbol{\beta}$ where each component is defined as $\beta_k = (-1)^k \lceil k/2 \rceil$ for $k = 1, \ldots, p$. This introduces a diverse range of predictor effects to test the robustness of the estimators.

ii) **Response Simulation**: For each observation i, we compute the linear predictor $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$. The response variable Y_i is then simulated from a univariate Gaussian distribution $\mathcal{N}(\eta_i, \sigma_i^2)$.

iii) **Heteroscedasticity and Normalisation**: To incorporate heteroscedasticity, the standard error σ_i is drawn from an absolute normal distribution with a mean of 10 and a standard deviation of 1. Finally, the response variable Y_i is normalised to ensure a zero mean across the entire dataset.

Step 3: Experimental Scale and Repetitions

i) **Dimensions and Sample Sizes**: The simulation considers varying numbers of covariates $m = p+1 \in \{2, 10, 25\}$. For each m, the sample size n is determined by specific ratios of the sample size to the number of covariates, namely $n/m \in \{10, 25, 50, 100\}$.

ii) **Statistical Reliability**: To ensure the statistical significance of the reported results, all quantities and performance metrics are computed based on $n = 1,000$ samples and $N = 1,000$ independent repetitions for each scenario.

C Description of WTI and Brent Data

This section provides additional technical details regarding the data sources, the rationale for selecting specific oil benchmarks, and the granular timeline of the structural segments used in the empirical analysis.

C.1 Market selection and data sources

The selection of WTI and Brent as our primary subjects is motivated by their roles as the two most significant global price benchmarks for crude oil. WTI serves as the underlying commodity for the *New York Mercantile Exchange (NYMEX)* futures and reflects North American supply–demand dynamics. Brent, traded on the *Intercontinental Exchange (ICE)*, acts as the benchmark for roughly two-thirds of the world's internationally traded physical crude oil. These benchmarks are standard in the commodity literature for analysing risk premiums, price discovery, and factor-based investment strategies (Sakkas and Tessaromatis, 2020; Bakshi et al., 2019).

The monthly data, spanning January 1980 to September 2024, were sourced from Bloomberg. This includes front-month futures prices used to calculate excess returns, a standard proxy for commodity investment performance in empirical finance (Yang, 2013), as well as various market- and liquidity-based metrics used to construct the predictive factors

(Hong and Yogo, 2012). While WTI data are robustly available from 1983 onwards, the inclusion of these specific liquidity metrics for Brent necessitated a start date of April 1998 to ensure a balanced panel of all covariates for the analysis.

C.2 Detailed factor background

The factors employed in our analysis are calculated based on methodologies established in the commodity investing literature, particularly those developed for identifying priced risk premia in futures markets (Sakkas and Tessaromatis, 2020). These factors capture distinct dimensions of the commodity risk premium and are described below.

Inventory and Term Structure: Factors such as *Basis* (slope of the term structure) and *Basis Momentum* (change in basis) capture the "roll yield" associated with the shape of the futures curve. The *Basis* factor is rooted in the Theory of Storage (Working, 1949) and the Hedging Pressure Hypothesis (Keynes, 1930), identifying backwardation and contango as fundamental drivers of expected returns. *Basis Momentum*, defined as the difference between the momentum signals of the first and second nearby contracts, provides compensation for commodity volatility and curve dynamics (Boons and Prado, 2019).

Trend and Sentiment: *Momentum* (past returns) captures the tendency of commodity returns to persist over a 12-month horizon (Miffre and Rallis, 2007). *Hedging Pressure* (commercial vs. non-commercial positioning) and *Open Interest* (market liquidity) reflect the positioning of commercial hedgers relative to speculators and the overall risk absorption capacity of the market (Hong and Yogo, 2012). Based on the theory of normal backwardation (Keynes, 1930), these factors proxy for the risk premium demanded by speculators for providing insurance to producers (Bessembinder, 1992). Due to data limitations, *Hedging Pressure* is omitted for Brent as the *Commodity Futures Trading Commission (CFTC)* reports primarily cover US-based exchanges like NYMEX.

Risk, Value, and Macro: *Skewness* (third moment of returns) and *Volatility* (return dispersion) address non-normal return distributions and "fat tails" typical of energy markets,

capturing compensation for jump and variance risks (Fernandez-Perez et al., 2018). *Inflation Beta* (sensitivity to inflation) captures the historically documented role of commodities as a hedge against unexpected rising price levels (Gorton and Rouwenhorst, 2006). Finally, the *Value* (spot-to-long-term-average ratio) factor identifies long-term mean reversion by comparing current prices to their five-year historical average (Asness et al., 2013).

C.3 Detailed structural breakdown of testing periods

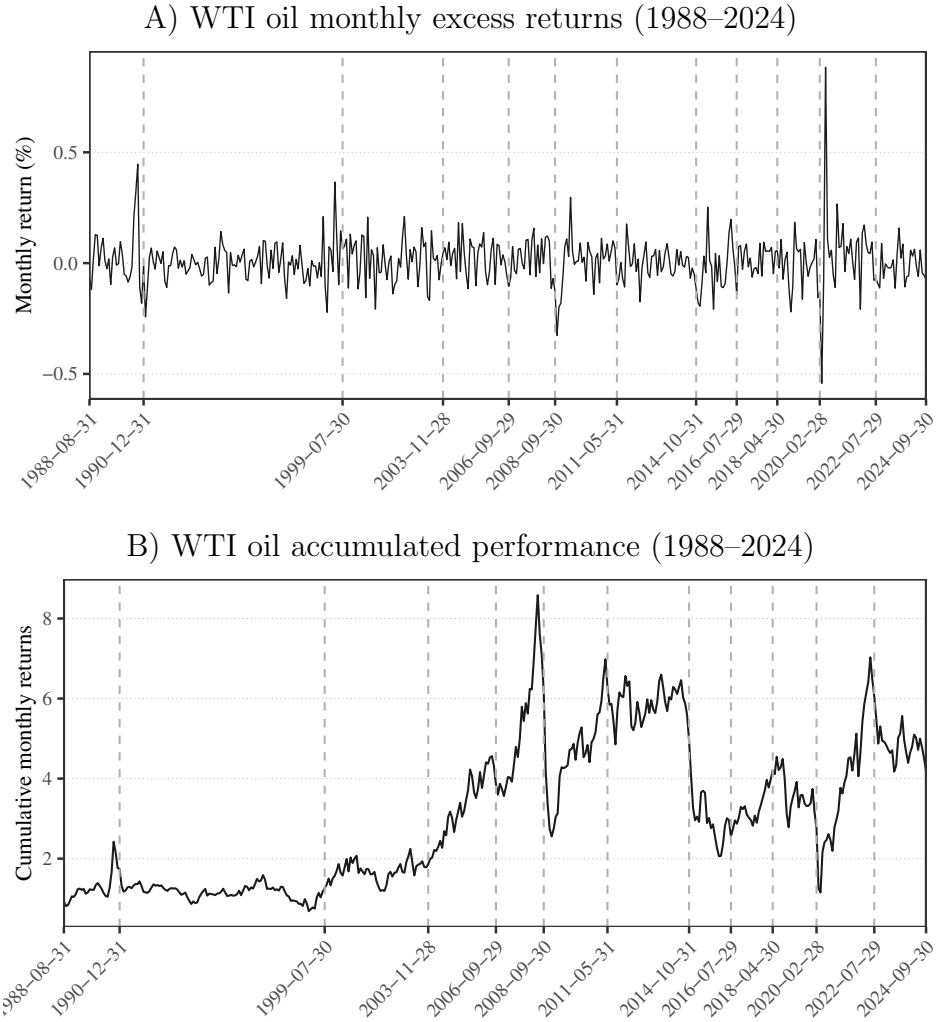
We applied the Bai and Perron (2003) test to the monthly returns to identify the structural breakpoints that define the OOS periods, as shown in Figures 2 and 3. This approach ensures that the segments align with genuine shifts in market volatility and mean-return levels. Table 5 summarises the 12 periods for WTI and 11 for Brent, including specific labels for notable market events beyond the primary crisis periods.

Table 5: **Detailed segmented periods for WTI and Brent with key market events**

	WTI Periods			**Brent Periods**		
Period	**Start**	**End**	**Key Market Event**	**Start**	**End**	**Key Market Event**
1	Aug 1988	Dec 1990	1990 Supply Shock	Apr 1998	Aug 2000	Asian Financial Crisis
2	Jan 1991	Oct 1993	Early 90s Oversupply	Sep 2000	Apr 2003	Early 2000s Recession
3	Nov 1993	Jan 1997	Mid-90s Expansion	May 2003	Aug 2005	China Demand Boom
4	Feb 1997	Aug 2000	Asian Financial Crisis	Sep 2005	Aug 2008	Pre-GFC Commodity Supercycle
5	Sep 2000	Aug 2008	2000s Commodity Supercycle	**Sep 2008**	**Jan 2011**	**Global Financial Crisis**
6	**Sep 2008**	**May 2011**	**Global Financial Crisis**	**Feb 2011**	**May 2014**	**Post-Crisis Recovery**
7	**Jun 2011**	**Jun 2014**	**Post-Crisis Recovery**	Jun 2014	Mar 2016	2014–16 Price Collapse
8	Jul 2014	Mar 2016	2014–16 Price Collapse	Apr 2016	Sep 2018	OPEC+ Production Cuts
9	Apr 2016	Sep 2018	OPEC+ Production Cuts	Oct 2018	Jan 2020	US–China Trade Tension
10	Oct 2018	Feb 2020	US–China Trade Tension	**Feb 2020**	**Jul 2022**	**COVID-19 Pandemic**
11	**Mar 2020**	**Jul 2022**	**COVID-19 Pandemic**	Aug 2022	Sep 2024	Post-Pandemic
12	Aug 2022	Sep 2024	Post-Pandemic	–	–	–

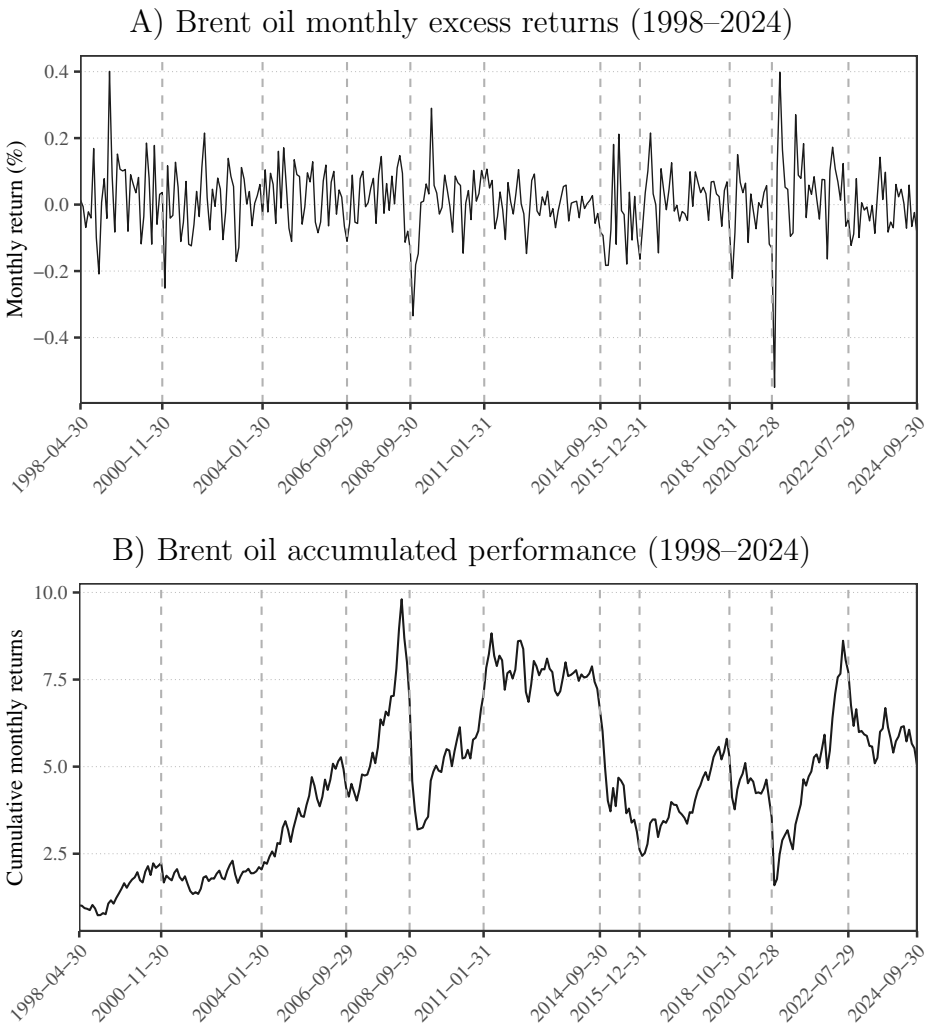
Notes. This table lists the segmented time periods identified by the Bai and Perron (2003) structural breakpoint test for both WTI and Brent crude oil datasets. Major global economic events impacting the volatility and price structure of both benchmarks are highlighted in bold. For each OOS iteration, models are trained on all data prior to the start of the current period to evaluate predictive accuracy in the subsequent regime.

Figure 2: Time series of WTI crude oil performance



Notes. This figure depicts the monthly excess returns (Panel A) and accumulated performance (Panel B) of WTI oil from August 1988 to September 2024. The solid red vertical lines indicate the dataset boundaries. The dashed orange vertical lines depict structural breakpoints identified by the Bai and Perron (2003) test, highlighting major economic or market events used for predictive segmentation.

Figure 3: Time series of Brent crude oil performance



Notes. This figure depicts the monthly excess returns (Panel A) and accumulated performance (Panel B) of Brent oil from April 1998 to September 2024. The solid red vertical lines indicate the dataset boundaries. The dashed orange vertical lines represent key structural breakpoints identified by the Bai and Perron (2003) test used to define periods for predictive modelling.

C.4 Regime characteristics and model stability

The identified structural periods encompass a wide range of volatility regimes, providing a rigorous test for both traditional and regularised estimators. As reflected in our empirical results, the relative performance of the estimators is highly contingent on the underlying macroeconomic environment. During periods characterised by stable expansion or moderate transitional dynamics—such as the Mid-90s Expansion (WTI Period 3), the 2000s Commodity Supercycle (WTI Period 5), and the Pre-GFC Commodity Supercycle (Brent Pe-

riod 4)—OLS frequently delivers highly competitive, and occasionally superior, predictive accuracy. In these low-to-moderate volatility regimes, historical factor correlations remain largely intact, allowing the unpenalised OLS estimator to efficiently capture the underlying data structure without the bias introduced by shrinkage methods.

By contrast, during extreme market regimes marked by structural shifts, OLS typically suffers from pronounced variance inflation. For instance, during the OPEC+ Production Cuts (WTI Period 9) and the COVID-19 pandemic (Brent Period 10), the breakdown of historical factor relationships leads OLS to produce substantial OOS forecast errors. While traditional shrinkage models—such as RR and Liu—provide essential cahnges to OLS under these conditions, they remain relatively sensitive to abrupt variance spikes. Because these methods rely on a global contraction controlled by a single penalty parameter, shocks to specific factors can result in either over-shrinkage of important signals or unstable, rapidly fluctuating parameter estimates. This sensitivity often undermines OOS predictive performance, as exemplified during the Global Financial Crisis (Brent Period 5), where RR and Liu underperform even OLS by overreacting to transient market noise.

In contrast, the PR framework, and particularly PR_c, preserves predictive stability by explicitly imposing a parity constraint, which serves as a structural safeguard. By focusing on a parity approach rather than raw parameter magnitude, PR is inherently less vulnerable to volatility spikes that can destabilise conventional penalised models. Unlike RR or Liu, which shrink parameters, the PR framework ensures that no single factor dominates the model's risk profile. This prevents overfitting to transient noise or extreme events, such as the severe negative WTI returns in April 2020 (WTI Period 11) or the 2014–2016 Price Collapse (Brent Period 7). Such structural regularisation is particularly valuable in non-stationary, highly volatile regimes, as the parity mechanism provides a stable budgeting bound that preserves predictive accuracy even when traditional correlation structures abruptly fail.