TI 2022-066/III
Tinbergen Institute Discussion Paper

Robust Observation-Driven Models Using Proximal-Parameter Updates

Revision: December 2022

Rutger-Jan Lange[1]
Bram van Os[1]
Dick van Dijk[1]

[1] Erasmus University Rotterdam

Robust Observation-Driven Models Using Proximal-Parameter Updates[*]

Rutger-Jan Lange[†], Bram van Os[‡] and Dick van Dijk[§]

Econometric Institute, Erasmus University Rotterdam

December 20, 2022

Abstract

We propose an observation-driven modelling framework that permits time variation in the model's parameters using a proximal-parameter (ProPar) update. ProPar maximizes the observation log-density with respect to the parameter vector, while penalizing the weighted ℓ_2 norm relative to the one-step-ahead prediction. This yields an *implicit* stochastic-gradient update; taking instead the explicit version would produce the popular class of score-driven models. For log-concave observation densities (even when misspecified), ProPar's robustness is evident from its muted response to outliers, stability under poorly specified learning rates, and global contractivity towards a pseudo-truth. We illustrate ProPar's usefulness for estimating time-varying regressions, volatility, and quantiles.

Keywords: Implicit gradient, Proximal point method, Robust filters, Score-driven methods, Time-varying parameter models.

[*] We gratefully acknowledge comments and suggestions from Christian Brownlees, Janneke van Brummelen, Leopoldo Catania, Guillaume Chevillon, Ana Galvao, Gustavo Freire, Peter Hansen, Onno Kleen, Erik Kole, Andrew Patton and the participants of the 15th Meeting of the Netherlands Econometric Study Group (NESG 2022), the 28th International Conference on Computing in Economics and Finance (CEF 2022), the 14th Annual Conference of the Society for Financial Econonometrics (SoFiE 2022), the Aarhus Workshop in Econometrics, the online SNDE Workshop for Young Scholars and the Econometrics Seminars and Workshop Series at the University of Amsterdam. Any remaining errors are our own.

[†] lange@ese.eur.nl (corresponding author)

[‡] vanos@ese.eur.nl

[§] djvandijk@ese.eur.nl

1 Introduction

Ample empirical evidence suggests it is often too restrictive to assume that model parameters remain constant for prolonged periods of time. In economics and finance, parameters are often found to be regime dependent or subject to structural breaks (e.g., Stock and Watson, 1996). Parameters may also change more gradually, following no discernible pattern; particularly in this case, it is unclear how to update them after observing new data. Ex-post estimators can be constructed in specific cases; e.g., in ARCH-type models (see Teräsvirta, 2009, for an overview) volatility is made time-varying using the squared shock, which provides an unbiased ex-post proxy of the true variance. In general, however, such proxies may be difficult to derive, inefficient, or nonexistent.

We introduce a comprehensive framework that allows a model's parameters to be made time-varying in an observation-driven setting by means of proximal-parameter (ProPar) updates. The proposed ProPar filter contains alternating prediction and update steps analogous to those in Kalman's (1960) filter. The key component of our framework is the ProPar update step, which ensures that the update remains proximal (i.e., close) to the prediction. Specifically, the ProPar parameter update is the solution to an optimization problem that maximizes the log-likelihood contribution of the current observation subject to a weighted ℓ_2 penalty centered at the one-step-ahead prediction. This setup has the advantage of (a) fully exploiting the log-likelihood contribution of the most recent observation, while (b) regularizing (i.e., limiting) the amount by which the update deviates from the prediction. The weighted ℓ_2 penalty is controlled by a positive-definite penalty matrix, the inverse of which can be viewed as a learning-rate matrix.

The first-order condition corresponding to the optimization problem solved in the ProPar update step can be formulated as an implicit stochastic-gradient update: implicit because the gradient is evaluated in the updated rather than the predicted parameter, and stochastic because it uses noisy data. In the optimization literature, such methods are known as proximal-point or proximal-gradient methods, and are recognized as inherently more stable than their explicit counterparts. Unlike explicit gradient updates, implicit approaches are guaranteed to improve the objective function—in our case, the log-likelihood contribution of the most recent observation. Explicit gradient methods are nonetheless widely used in observation-driven models, e.g., in the popular class of score-driven models (see further section 1.1).

The ProPar filter has several attractive theoretical properties. As outlined below and demonstrated at length in the paper, these properties are typically sought in observation-driven models, but rarely combined in a single framework; indeed, we are unaware of other approaches offering the same combination of advantages.

1. The ProPar filter is invertible (Theorem 1) under mild restrictions and assuming concavity of the researcher-postulated logarithmic density; i.e., any differences stemming from the initialization of the filter disappear almost surely and exponentially fast. This holds even in the near absence of assumptions placed on the data-generating process; as such, this result is highly robust to model misspecification.

2. The ProPar update yields an updated density that improves on the predicted density (even if both are misspecified) by reducing the local Kullback-Leibler divergence relative to the true density (Proposition 3). This provides a local information-theoretic foundation in the spirit of Blasques et al. (2015), but it is substantially stronger as we are able to consider sizable (i.e., non-infinitesimal) adjustments of the time-varying parameters.

3. The ProPar update is globally contractive (Theorem 2) in expectation to a small region around the (pseudo-)truth. This means that, on average, the update is more accurate than the prediction on which it is based, with the largest improvements expected for the worst predictions. This contraction is global in that the predictions may be arbitrarily bad; it is also robust in that it holds for an arbitrary (positive-definite) learning-rate matrix. Only when the prediction is very close to the (pseudo-)true parameter may the update be less accurate; this is no limitation but unavoidable when using noisy data.

Beyond these theoretical advantages, the ProPar framework also has practical benefits. It automatically coordinates the update of multiple interacting parameters. Additionally, parameter constraints may be incorporated without necessarily requiring parameter transformations (e.g., link functions).

We evaluate the empirical performance of the ProPar filter in three different settings, demonstrating the theoretical and practical advantages outlined above. First, we consider a linear regression of daily Microsoft equity returns on the market factor, where the regression coefficient (i.e., the slope) is made time-varying. Second, we model time-varying volatility using daily S&P500 data. In these illustrations, the robustness of the ProPar filter successfully mitigates the effects of large shocks, making it better able to cope with rare events compared to standard alternatives. Third, we consider growth-at-risk estimates captured by the lower quantiles of quarterly US GDP growth. The corresponding ProPar update yields an implicit version of Engle and Manganelli's (2004) adaptive CAViaR model, where ProPar has the advantage in that its quantile update cannot be more extreme than the observation just received. This enhanced stability, together with simple parameter restrictions, ensures that simultaneously modeled quantiles remain properly ordered, thus avoiding the quantile-crossing problem faced by other methods.

In Section 2, we outline the ProPar methodology, highlight the differences with conventional score-driven models, and illustrate the proposed method by way of an example. Section 3 presents the theoretical properties, focusing on stability, local information-theoretic optimality, and global contractivity, while Section 4 discusses maximum-likelihood estimation of the static parameters. Section 5 contains the empirical illustrations. Finally, Section 6 concludes with implications and recommendations. All proofs are provided in the online Appendix.

1.1 Related literature

This paper ties in with two main strands of literature. First, the ProPar filter can be viewed as a stochastic version of Rockafellar's (1976) proximal-point algorithm, which combines a function to be optimized with a quadratic penalty involving some previous iterate. As our log-likelihood function involves (random) observations drawn from the true density, the ProPar filter can be considered a stochastic proximal-point method (e.g., Ryu and Boyd, 2016; Bianchi, 2016; Patrascu and Necoara, 2018; Asi and Duchi, 2019). The first-order condition associated with the proximal optimization can also be rewritten as an implicit stochastic-gradient step (e.g., Toulis and Airoldi, 2015; Toulis et al., 2016; Toulis and Airoldi, 2017; Toulis et al., 2021). As in the optimization literature, we recognize the advantage of implicit over explicit gradient methods in terms of enhanced stability. A key difference is, however, that we consider a setup in which the parameter to be estimated is not constant but changing over time—i.e., we are interested in tracking a moving target.

Second, this article is related to observation-driven time-varying parameter models, so labeled in Cox et al. (1981). One benefit of these models is that the likelihood can be computed in closed form, enabling numerical maximum-likelihood (ML) estimation. Within this class of model, dynamic conditional score (DCS; Harvey, 2013) models and generalized autoregressive score (GAS; Creal et al., 2013) models, as they are variously known, use the score of the log-likelihood function to propagate time-varying parameters. This score-driven framework encompasses many established models, such as the GARCH model, and is popular for its ease of use and strong forecasting performance (e.g., Creal et al., 2014; Harvey and Luati, 2014; Koopman et al., 2016; Harvey and Lange, 2017; Opschoor et al., 2018; Gorgi, 2020); some 300 articles using this method are listed on www.gasmodel.com.

As the current article demonstrates, this class of score-driven models can be obtained within the ProPar framework by replacing, at every time step, the logarithmic observation density by its local-linear approximation around the one-step-ahead prediction. Under this approximation, ProPar's implicit stochastic-gradient update is replaced by its explicit version, yielding standard (i.e., explicit) score-driven models. Avoiding the local-linear approximation and preserving the full logarithmic observation density enables us to generalize several attractive properties of score-driven models from the local to the global setting. We are able to derive a stronger form of local information-theoretic optimality than is available for score-driven models (Blasques et al., 2015); e.g., we need not restrict ourselves to arbitrarily small adjustments in the time-varying parameter. Our invertibility results are stronger, placing fewer demands on the data-generating process, and our framework allows for global contractivity towards a (pseudo-)true parameter, a property with no obvious equivalent in the literature on explicit score-driven models.

2 Proximal-parameter framework

2.1 Prediction-update recursion

We consider an $N \times 1$ variable of interest y_t, observed at times $t = 1, \dots, T$, drawn from an observation density $p_0(\cdot|\theta_t^0, \psi^0, \mathcal{F}_{t-1})$, where θ_t^0 is a time-varying parameter vector taking its values in some parameter space Θ^0, ψ^0 is a vector of static shape parameters, and $\mathcal{F}_{t-1}$ denotes the information set at time $t-1$, thereby permitting the dependence on exogenous variables and/or lags of y_t. For readability, the dependence on ψ^0 and $\mathcal{F}_{t-1}$ is henceforth suppressed.

The aim of this paper is to devise a modeling framework that attempts to approximate the true distribution $p_0(\cdot|\theta_t^0)$. To this end, we propose a filter that alternates between prediction and update steps. Specifically, let $p(\cdot|\theta_t)$ denote the researcher-postulated density, which may or may not be correctly specified, where θ_t denotes a $K \times 1$ vector of time-varying parameters that can take values in some non-empty convex parameter space $\Theta \subseteq \mathbb{R}^K$. As above, any additional dependence on static shape parameters and/or other information available at time $t-1$ is permitted but suppressed for readability. We denote the predicted and updated parameter vectors by $\theta_{t|t-1} \in \Theta$ and $\theta_{t|t} \in \Theta$, which reflect the researcher's estimates of θ_t using the information set available at times $t-1$ and t, respectively.

The main difficulty in working with time-varying parameter models lies in specifying how $\theta_{t|t}$ should differ from $\theta_{t|t-1}$ after observing y_t. We argue that a sound update scheme should satisfy at least two natural criteria. First, the update should be in accordance with the likelihood, such that $p(y_t|\theta_{t|t}) \geq p(y_t|\theta_{t|t-1})$: i.e., the updated parameter yields an improved fit when evaluated at the observed data y_t. Second, as each observation y_t is inherently noisy, it is desirable to regularize the amount by which the update $\theta_{t|t}$ deviates from the prediction $\theta_{t|t-1}$. Penalizing the magnitude of $\theta_{t|t} - \theta_{t|t-1}$ prohibits the filter from becoming excessively volatile.

To satisfy both criteria, this article proposes the class of proximal-parameter (ProPar) models. These models perform the parameter update at time t by maximizing the researcher-postulated logarithmic observation density $\log p(y_t|\cdot)$ subject to a weighted ℓ_2 penalty centered at the prediction $\theta_{t|t-1}$. That is, we consider the parameter update

$$\theta_{t|t} \;:=\; \underset{\theta\in\Theta}{\operatorname{argmax}}\; f(\theta|y_t, \theta_{t|t-1}, P_t), \tag{1}$$

where

$$f(\theta|y_t, \theta_{t|t-1}, P_t) \;:=\; \log p(y_t|\theta) \;-\; \frac{1}{2}\Big\|\theta - \theta_{t|t-1}\Big\|^2_{P_t}. \tag{2}$$

Here, $f(\theta|y_t, \theta_{t|t-1}, P_t)$ denotes the "regularized" log-likelihood contribution and $\|x\|^2_{P_t} = x'P_t x$ is the squared ℓ_2 norm with respect to a $K \times K$ positive-definite penalty matrix P_t. By formulating the parameter update as the solution to a maximization problem, the proposed method has several favorable characteristics. First, all information in the conditional density is utilized to update the parameter, as opposed to, e.g., moment information only. Second, elements of the parameter update $\theta_{t|t}$ are automatically interdependent, because jointly they represent the solution to the multivariate optimization problem (1). Third, the update $\theta_{t|t}$ is automatically contained in the correct space Θ and does not necessarily require a link function to be specified (although we may employ link functions for other reasons). We may constrain Θ to any non-empty convex subset, allowing for straightforward incorporation of a great variety of convex and possibly non-differentiable constraints.

The weighted ℓ_2 penalization yields tractable updates and can be interpreted as a second-order Taylor expansion around $\theta_{t|t-1}$ of an arbitrary, but more complicated, loss function, where P_t acts as the corresponding Hessian. Furthermore, the update in the ProPar approach defined in equations (1) and (2) takes a comparable form to Rockafellar's (1976) classic proximal-point algorithm, which similarly considers the optimization of a target function—in our case, the log-likelihood contribution of the observation y_t—subject to a quadratic penalty. Because the likelihood contribution is based on the (a priori random) realization y_t, our approach can be viewed as a stochastic proximal-point method (Asi and Duchi, 2019). Thanks to their favorable characteristics, proximal-point methods are widely employed in optimization (see Section 1.1). We will exploit these characteristics to obtain a variety of attractive properties of the ProPar filter in terms of stability and optimality (see Section 3).

Below, we formalize two standard assumptions (Assumptions 1 and 2) regarding the existence and uniqueness of the solution to the maximization problem (1). We add two further assumptions (Assumptions 3 and 4) that allow us to characterize its solution using a standard first-order condition. While this simplification is not absolutely necessary (e.g., we could work with subgradients), it benefits the clarity of exposition and improves mathematical tractability.

Assumption 1 (Existence) *The solution set of* $\underset{\theta \in \Theta}{\operatorname{argmax}} f(\theta|y_t, \theta_{t|t-1}, P_t)$ *is non-empty with probability one.*

Assumption 2 (Strictly concave regularized log likelihood) $f(\theta|y_t, \theta_{t|t-1}, P_t)$ *is proper strictly concave in* θ*,* $\forall \theta \in \Theta$ *with probability one.*

Assumption 3 (Interior solution) $\theta_{t|t} \in \operatorname{Int}(\Theta)$ *with probability one.*

Assumption 4 (Differentiability) $\log p(y_t|\theta)$ *is at least once continuously differentiable in* θ*,* $\forall \theta \in \operatorname{Int}(\Theta)$ *with probability one.*

Under Assumptions 1 through 4, the first-order condition for the parameter update $\theta_{t|t}$ in the maximization problem (1) can be rearranged as

$$\theta_{t|t} \; = \; \theta_{t|t-1} \, + \, H_t \, \nabla(y_t|\theta_{t|t}), \tag{3}$$

where the inverse penalty $H_t := P_t^{-1}$ is referred to the learning-rate matrix and $\nabla(y_t|\theta_{t|t}) := (\partial \log p(y_t|\theta)/\partial\theta)|_{\theta=\theta_{t|t}}$ is the score vector, both at time t. Representation (3) demonstrates that the ProPar framework yields a gradient-type parameter update. The learning-rate matrix H_t controls the step size and allows for different learning rates and interactions between the different time-varying parameters. Crucially, the score is evaluated at the update $\theta_{t|t}$ rather than the prediction $\theta_{t|t-1}$. This means that update (3) is an *implicit* gradient method; i.e., the parameter update $\theta_{t|t}$ appears on both sides of the equation, hence is not immediately computable. Because the update $\theta_{t|t}$ is also stochastic—it is based on the a priori random realization y_t—our framework is closely related to implicit stochastic-gradient methods (see section 1.1). While the first-order condition (3) may not allow a closed-form solution, Assumptions 2 and 4 guarantee that the global solution to optimization problem (1) can always be found numerically using standard optimization techniques (e.g., quasi-Newton methods).

An attractive property of implicit updates is their enhanced stability and optimality relative to explicit gradient methods, which use the gradient evaluated in the prediction $\theta_{t|t-1}$ rather than the update $\theta_{t|t}$. Implicit updates are guaranteed to increase the value of the objective function—in our case the log-likelihood contribution of y_t—whereas explicit versions may decrease the objective function value when the step size is too large; i.e., explicit methods may "overshoot". To mitigate this problem, explicit gradient methods must often be implemented with smaller learning rates. In contrast, when the objective function is strictly concave, implicit gradient methods can be shown to converge to the global optimum for *any* positive definite learning-rate matrix H_t (Toulis and Airoldi, 2017). For this reason, implicit optimization techniques are widely employed in statistics and machine learning (e.g., Kulis and Bartlett, 2010; Li et al., 2014).

A key difference with the existing literature on implicit gradient methods in optimization is that we consider a setting in which the true parameter is time-varying rather than constant. In the optimization literature, the learning-rate matrix H_t is typically set to be decreasing over time (e.g., $H_t = O(t^{-1})$), such that the parameter asymptotically converges to some constant pseudo-true value. Here we are interested in tracking a time-varying true parameter; hence, our filtered path must not converge over time, but remain responsive even asymptotically. To achieve this we may keep H_t constant over time, i.e., set $H_t = H$ for all t, where H may contain static parameters that are to be estimated (see Section 4).

To complete our dynamic setup, the ProPar update step (3) is complemented with a prediction step that generates one-step-ahead forecasts. For simplicity, we consider a linear first-order specification as follows:

$$\theta_{t+1|t} \;=\; \omega \;+\; \Phi\,\theta_{t|t}, \tag{4}$$

where ω is a $K \times 1$ vector of constants and Φ is a $K \times K$ autoregressive matrix. Conditions ensuring stable recursions are discussed in the next section. The requirement $\theta_{t+1|t} \in \Theta$ can typically be fulfilled by appropriate parameter restrictions and/or link functions. In principle, the prediction step (4) could be generalized to allow for non-linear and/or higher-order dynamics. However, as no additional information is available during the prediction step, a more complicated structure may not yield immediate benefits. For simplicity, do not pursue this here.

To sum up, suppose we are given (a) some data $\{y_t\}$ for $t = 1, 2, \ldots, T$, (b) a researcher-postulated density $p(\cdot|\theta)$ satisfying Assumptions 1 through 4, (c) a set of prediction parameters ω and Φ, (d) a sequence of penalization matrices $\{P_t\}$, and (e) some initial estimate $\theta_{0|0} \in \Theta$. Then we can iteratively apply the prediction-update recursion consisting of the prediction step (4) and the update step (1) or, equivalently, (3). Together, these recursions produce sequences of parameter predictions, $\{\theta_{t|t-1}\}$, and parameter updates, $\{\theta_{t|t}\}$, such that the description of the ProPar filter is now complete.

2.2 Relationship with (explicit) score-driven filters

The implicit gradient-type update (3) suggests a close connection with the large literature on score-driven models. Here we show that linearizing the logarithmic observation density in the optimization problem (1) produces the familiar explicit gradient update, denoted by $\theta^{\mathrm{e}}_{t|t}$. Specifically, suppose we approximate the logarithmic observation density in equation (2) using a first-order Taylor expansion around the prediction $\theta_{t|t-1}$, i.e., $\log p(y_t|\theta) \approx \log p(y_t|\theta_{t|t-1}) + \langle \theta - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1})\rangle$, where $\langle x_1, x_2\rangle := x_1' x_2$ denotes the inner product. To avoid boundary solutions, we suppose the maximization is over the Euclidean space $\mathbb{R}^K$. Because the regularized log-likelihood contribution $f(\cdot|\cdot,\cdot,\cdot)$ in optimisation problem (1) now contains a linear target in combination with a quadratic penalty, the optimization can be performed in closed form. Indeed, the resulting linearized version of optimization (1) and associated first-order condition now read

$$\theta^{\mathrm{e}}_{t|t} := \underset{\theta\in\mathbb{R}^K}{\operatorname{argmax}} \left\{ \log p(y_t|\theta_{t|t-1}) + \langle \theta - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1})\rangle - \frac{1}{2}\|\theta - \theta_{t|t-1}\|^2_{P_t} \right\}, \tag{5}$$

$$\theta^{\mathrm{e}}_{t|t} \;=\; \theta_{t|t-1} \,+\, H_t\,\nabla(y_t|\theta_{t|t-1}). \tag{6}$$

The score on the right-hand-side of the explicit update (6) is evaluated at the prediction $\theta_{t|t-1}$ rather than the update $\theta_{t|t}$, such that $\theta^{\mathrm{e}}_{t|t}$ is immediately computable. In combination with prediction step (4), the explicit updating strategy (5) yields a well-known class of explicit score-driven models, known either as dynamic conditional score (DCS) models (Harvey, 2013) or generalized autoregressive-score (GAS) models (Creal et al., 2013). This class of score-driven models can be regarded as a first-order approximation to the implicit update (3), similar to how explicit gradient methods in optimization are viewed as first-order approximations of implicit- or proximal-gradient methods.

It is natural to ask whether the implicit and explicit update strategies yield similar results. Proposition 1 below shows that both strategies suggest adjustments of the time-varying parameter that point roughly in the same direction. Geometrically, the angle between the difference vector $\theta_{t|t} - \theta_{t|t-1}$ and the explicit version of the score (i.e., $\nabla(y_t|\theta_{t|t-1})$) cannot exceed 90 degrees.

Proposition 1 (Gradient alignment) *Fix $t > 0$ and let Assumptions 1 and 2 hold. Let a prediction $\theta_{t|t-1} \in \Theta$ and positive-definite penalty $P_t \in \mathbb{R}^{K\times K}$ be given and assume that $\nabla(y_t|\theta_{t|t-1})$ is well-defined. Compute $\theta_{t|t}$ by the update step* (1). *Then, with probability one,*

$$\langle \theta_{t|t} - \theta_{t|t-1}, \nabla(y_t|\theta_{t|t-1})\rangle \;\geq\; 0. \tag{7}$$

For a scalar time-varying parameter (i.e., $K = 1$), Proposition 1 implies that $\theta^{\mathrm{e}}_{t|t} - \theta_{t|t-1}$ and $\theta_{t|t} - \theta_{t|t-1}$ have the same sign; in this case, the implicit score is "score equivalent", using the definition of Blasques et al. (2015).

Corollary 1 (Gradient-sign concordance in one dimension) *Fix $t > 0$ and let Assumptions 1 to 4 hold. Let a prediction $\theta_{t|t-1} \in \Theta \subseteq \mathbb{R}$ and penalty $P_t > 0$ be given. Compute $\theta_{t|t}$ by the update step* (1)*. Then, with probability one,*

$$\operatorname{sign}(\nabla(y_t|\theta_{t|t})) \; = \; \operatorname{sign}(\nabla(y_t|\theta_{t|t-1})). \tag{8}$$

To say more about the properties of the ProPar update step (1), we require more information regarding the shape of the log-likelihood function $\log p(y_t|\theta)$. In this paper, we focus on the family of concave log-likelihood functions, which allows us to derive a set of particularly strong optimality and stability properties.

Assumption 5 (Log-concave observation density) $\log p(y_t|\theta) + \alpha_t/2 \, \|\theta\|^2$ *is concave in θ for some $\alpha_t \geq 0$, $\forall \theta \in \Theta$, with probability one.*

Assumption 5 is a stronger version of Assumption 2, as we now impose concavity on the log-likelihood contribution itself, rather than on its regularized version (2). The strength of concavity is measured by $\alpha_t \geq 0$, where the boundary case $\alpha_t = 0$ implies concavity while $\alpha_t > 0$ implies α_t-strong concavity. A large collection of popular logarithmic densities, as illustrated in the empirical section, are concave in their parameters. While Assumption 5 yields strong theoretical results, the optimization literature suggests that implicit gradient methods remain effective in practice if the logarithmic density fails to be concave (e.g., Hare and Sagastizábal, 2009); in fact, the global nature of the proximal update (1) may further enhance the advantages relative to explicit methods in such a setting (e.g., Grimmer et al., 2022, p. 31).

It is well known in the optimization literature that the implicit gradient update is a "shrunken" version of the explicit gradient update (e.g., Toulis and Airoldi, 2015). Proposition 2 reflects this relationship in our setting.

Proposition 2 (Step-size shrinkage) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Let a prediction $\theta_{t|t-1} \in \Theta$ and positive-definite penalty $P_t \in \mathbb{R}^{K \times K}$ be given. Based on the observation y_t, compute $\theta_{t|t}$ by the implicit update* (3) *and $\theta^{\mathrm{e}}_{t|t}$ by the explicit update* (6)*. Then, with probability one,*

$$\left\|\theta_{t|t} - \theta_{t|t-1}\right\|^2_{P_t + 2\alpha_t I_K} \; \leq \; \left\|\theta^{\mathrm{e}}_{t|t} - \theta_{t|t-1}\right\|^2_{P_t}, \tag{9}$$

where I_K is the identity matrix of size K.

Inequality (9) features a weighted norm on both sides, where the weight matrix on the left-hand-side has a diagonal that is increased by a multiple of the identity matrix. As a result, the vector inside the norm on the left-hand-side must be smaller in some sense than the vector inside the norm on the right-hand side. The magnitude of the shrinkage depends on the ratio between the strength of concavity α_t and the penalty P_t, where a larger α_t or smaller P_t imply more shrinkage. In the scalar case (i.e., $K = 1$), equation (9) can be written as $\|\theta_{t|t} - \theta_{t|t-1}\|^2 \leq \frac{P_t}{P_t+2\alpha_t}\|\theta^{\mathrm{e}}_{t|t} - \theta_{t|t-1}\|^2$, where $P_t/(P_t + 2\alpha_t) \in (0, 1]$ is the shrinkage factor.

In practice, the shrinkage of the vector $\theta_{t|t} - \theta_{t|t-1}$ evident from equation (9) provides an additional level of robustness that is particularly useful for dealing with outliers. In the presence of outliers, the learning-rate matrix H_t must typically be reduced in magnitude to ensure that the filter is not excessively impacted by such aberrant observations. The shrinkage property (9) mitigates this problem, enabling ProPar models to use larger learning rates relative to standard (i.e., explicit) score-driven models. In the optimization literature, the fact that implicit strategies often allow for larger learning rates is well known (e.g., Toulis and Airoldi, 2017).

2.3 Example: Linear regression with time-varying slopes

This section illustrates several attractive properties of the ProPar framework using a linear regression model with time-varying parameters.

Example 1 (Linear regression) *Consider a linear regression model with dependent variable $y_t \in \mathbb{R}$ and independent variable $x_t \in \mathbb{R}^K$, i.e.*

$$y_t \;=\; \beta_t' x_t + \varepsilon_t, \qquad \varepsilon_t \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2), \tag{10}$$

where β_t is a $K \times 1$ vector of time-varying parameters and ε_t is an i.i.d. normally distributed innovation with variance σ^2. Then the ProPar update (1) can be computed in closed form (see Appendix A for details) as

$$\beta_{t|t} \;=\; \beta_{t|t-1} + \frac{\sigma^2}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \, \nabla(y_t|\beta_{t|t-1}, x_t), \tag{11}$$

where $H_t = P_t^{-1}$ is the learning-rate matrix and $\nabla(y_t|\beta_{t|t-1}, x_t)$ denotes the explicit score given as

$$\nabla(y_t|\beta_{t|t-1}, x_t) \;=\; \frac{y_t - \beta'_{t|t-1} x_t}{\sigma^2} x_t. \tag{12}$$

Example 1 illustrates the shrinkage result of Proposition 2 for the linear regression model. The right-hand-side of equation (11) features the shrinkage factor $\sigma^2/(\sigma^2 + \|x_t\|^2_{H_t}) \in (0, 1]$, which would be absent (i.e., equal to unity) in the case of an explicit score-driven update; hence, update (11) can be viewed as a robustified version of the explicit score-driven update. The amount of shrinkage is increasing in the magnitude of the explanatory variable (i.e., $\|x_t\|^2_{H_t}$) and decreasing in the observation variance (i.e., σ^2). For the ProPar update (11), it is easy to show that if a particular element of x_t tends to infinity in an absolute sense (i.e., $|x_{i,t}| \to \infty$ for some i), then the corresponding element of $\beta_{t|t}$ goes to zero (i.e., $\beta_{i,t|t} \to 0$), while the other elements remain unchanged at their predicted values (i.e., $\beta_{j,t|t} \to \beta_{j,t|t-1}$ for $j \neq i$). The fact that the shrinkage factor depends on the realization of the exogenous variable x_t appears to be distinctive for the ProPar version of the model; i.e., we are unaware of (explicit) score-driven models with this property.

Another difference with explicit score-driven models is that the ProPar update (11) remains bounded as the learning-rate matrix H_t grows larger (i.e., in a positive definite sense). This can be seen by noting that H_t appears not only in front of the score, but also in the denominator of the shrinkage factor. The practical relevance of this observation is that the ProPar filter is robust against the (suboptimal) choice of the learning rate, whereas explicit score-driven models tend to require more careful finetuning.

3 Theory

3.1 Stability

Turning to the stability properties of the proposed framework, we are particularly interested in providing sufficient conditions for filter invertibility, meaning that filtered paths based on identical data but with different initializations convergence exponentially fast over time. First, we show in Lemma 1 that the update step (1) admits strong contraction properties under Assumptions 1 through 5. We note that no additional conditions are imposed on the true data-generating process (DGP), as discussed in further detail below.

Lemma 1 (Prediction-to-update stability) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Let $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ denote two predictions in Θ, which are combined with the observation y_t in the update step (1) to yield two corresponding parameter updates, $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$, respectively. Then, with probability one,*

$$\left\|\theta_{t|t} - \tilde{\theta}_{t|t}\right\|^2_{P_t + 2\alpha_t I_K} \;\leq\; \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|^2_{P_t}. \tag{13}$$

If, in addition, the postulated log-likelihood function $\theta \mapsto \log p(\cdot|\theta)$ is twice differentiable, then, with probability one, the Jacobian matrix $\frac{\partial \theta_{t|t}}{\partial \theta'_{t|t-1}}$ has all eigenvalues in $(0,1]$. This interval becomes $(0,1)$ when $\alpha_t > 0$.

The first part of Lemma 1 indicates that the update step of the ProPar filter is non-expansive in the norm $\|\cdot\|_{P_t}$, i.e., the update step does not magnify (and possibly shrinks) the distance between different paths. The second part of Lemma 1 shows that the eigenvalues of the Jacobian $\partial\theta_{t|t}/\partial\theta'_{t|t-1}$ are in the unit interval if the log-likelihood function is twice continuously differentiable, which reflects an alternative definition of non-expansiveness. For a strongly concave log-likelihood function (i.e., $\alpha_t > 0$), we obtain a strict contraction in the norm $\|\cdot\|_{P_t}$ as long as the predictions are not identical (i.e., $\theta_{t|t-1} \neq \tilde{\theta}_{t|t-1}$). In this case, the eigenvalues of the Jacobian are strictly bounded between zero and one. The strength of the contraction is determined by the strength of concavity α_t and the penalty matrix P_t. Interestingly, Lemma 1 does not require further assumptions on the DGP.

To obtain a strictly contracting prediction-to-prediction mapping from time t to time $t+1$, it is sufficient to have both the update and prediction steps be non-expansive in the norm $\|\cdot\|_{P_t}$ with at least one of them being strictly contractive. That is, when $\alpha_t = 0$, the prediction mapping from $\theta_{t|t}$ to $\theta_{t+1|t}$ must be strictly contracting in the norm $\|\cdot\|_{P_t}$. When $\alpha_t > 0$, on the other hand, it is sufficient for the prediction step to be non-expansive. For example, the identity mapping $\theta_{t+1|t} = \theta_{t|t}$ is non-expansive and often useful in practice.

A sufficient condition for non-expansiveness (contractiveness) of the prediction step in the norm $\|\cdot\|_{P_t}$ is that $P_t \succeq \Phi' P_t \Phi$ ($P_t \succ \Phi' P_t \Phi$). Here, the notation $X \succeq Y$ ($X \succ Y$) indicates that $X - Y$ has non-negative (strictly positive) eigenvalues for two symmetric real-valued matrices X and Y of the same size. This requirement is equivalent to $\|\Phi\|_{P_t} \leq 1$ ($\|\Phi\|_{P_t} < 1$), where $\|X\|_{P_t}$ is the induced operator norm of a matrix $X \in \mathbb{R}^{K\times K}$. This condition is closely related to the discrete Lyapunov equation (e.g., Anderson and Moore, 2012). Lemma 2 summarizes the contraction of the prediction-to-prediction mapping in the norm $\|\cdot\|_{P_t}$.

Lemma 2 (Prediction-to-prediction stability) *Fix $t > 0$ and let Assumptions 1 to 5 hold. Let P_t be given with $P_t \succeq \Phi' P_t \Phi$. Let $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ denote two predictions in Θ that are used in the update step (1) to yield two corresponding parameter updates, $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$, and subsequently passed to the prediction step (4) to yield two predictions, $\theta_{t+1|t}$ and $\tilde{\theta}_{t+1|t}$. Then, with probability one,*

$$\left\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\right\|^2_{P_t} \;\leq\; \kappa_t \left\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\right\|^2_{P_t}, \tag{14}$$

where the contraction coefficient κ_t is

$$\kappa_t \;=\; \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t) + 2\alpha_t}, \tag{15}$$

where $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote the largest and smallest eigenvalues of X. If either $\alpha_t > 0$ or $P_t \succ \Phi' P_t \Phi$, then, with probability one, $\kappa_t \in [0, 1)$.

The strength of the contraction of the prediction-to-prediction mapping at time t is measured by κ_t, which is a function of the strength of concavity α_t, the penalty P_t and the autoregressive matrix Φ. For a scalar time-varying parameter, the standard condition $|\Phi| < 1$ is sufficient to yield $\kappa_t \in [0, 1)$. In the multiple-parameter setting, $\Phi'\Phi \prec I_K$ implies $\Phi' P_t \Phi \prec P_t$ when (a) Φ and P_t are both diagonal or (b) either Φ or P_t is a constant multiple of the identity. In this case, the standard condition that the spectral norm of Φ should be less than one is sufficient to yield $\kappa_t \in [0, 1)$. To allow for more richly parameterized Φ and P_t, we could allow Φ to be time-varying by expressing it in terms of P_t as

$$\Phi_t = P_t^{-1/2} V P_t^{1/2}, \tag{16}$$

where V is a $K \times K$ matrix of static autoregressive parameters with $\|V\|_2 < 1$, where $\|\cdot\|_2$ denotes the ℓ_2 operator norm. It is straightforward to show that this transformation implies $P_t \succ \Phi_t' P_t \Phi_t$. The matrices Φ_t and V are then similar; i.e., they have the same eigenvalues. Alternatively, $P_t = P$ could be taken to be constant for all t and expressed as the solution to (the discrete version of) Lyapunov's equation $P - \Phi' P \Phi = \Delta \succ 0$, which has a unique solution $P \succ 0$ parameterized in terms of Φ and $\Delta \succ 0$. The strict inequalities in this entire paragraph are permitted to become weak if we additionally require $\alpha_t > 0$.

For the effects of the initialization to disappear exponentially fast, it is required that the composition of all prediction-to-prediction mappings is contractive. A sufficient (but stronger than necessary) condition is that each individual prediction-to-prediction mapping is contractive in a single norm that is the same (i.e., shared) across all mappings over time. Theorem 1 formulates sufficient conditions for the existence of such a shared norm and contains an invertibility result that is crucial in enabling maximum-likelihood estimation of the static parameters (e.g., Straumann and Mikosch, 2006). This desirable invertibility property also ensures that numerical errors do not accumulate during implementation in practice, a concern also expressed for the Kalman filter (Anderson and Moore, 2012).

Theorem 1 (Invertibility) *For all $t > 0$, let Assumptions 1 to 5 hold, with additionally either (a) $P_t \succ \Phi' P_t \Phi$ or (b) $P_t \succeq \Phi' P_t \Phi$ and $\alpha_t > 0$. In addition, let there be some*

$\bar{P}, A \in \mathbb{R}^{K\times K}$ *with* $\bar{P} \succ A \succ O_{K\times K}$ *and a sequence* $\{\rho_t > 0\}$ *such that for all* $t > 0$*, with probability one,*

$$\kappa_t P_t \,+\, \rho_t A \;\preceq\; \rho_t\,\bar{P} \;\preceq\; P_t, \tag{17}$$

where κ_t *is defined in (15). Take two initial values* $\theta_{0|0} \in \Theta$ *and* $\tilde{\theta}_{0|0} \in \Theta$*, yielding two sequences* $\{\theta_{t|t-1}\}$ *and* $\{\tilde{\theta}_{t|t-1}\}$*, respectively. Then the filter composed of (1) and (4) is* invertible*, i.e., there exists a constant* $c_{(\cdot)} > 1$ *such that as* $t \to \infty$*, with probability one,*

$$\lim_{t\to\infty} c_{(\cdot)}^{t} \left\| \theta_{t|t-1} - \tilde{\theta}_{t|t-1} \right\|_{(\cdot)}^{2} \;\to\; 0, \tag{18}$$

for any norm $(\cdot)$*.*

Equation (17) in Theorem 1 expresses a sufficient condition for a contraction of all prediction-to-prediction mappings in the common norm $\|\cdot\|_{\bar{P}}$, where $\bar{P}$ is a constant matrix satisfying inequality (17). For a scalar time-varying parameter, this condition is guaranteed irrespective of the sequence $\{P_t\}$ whenever the standard condition $|\Phi| < 1$ holds. For the unit-root case $|\Phi| = 1$, it is sufficient that $\{P_t\}$ is upper bounded while $\{\alpha_t\}$ is strictly lower bounded away from zero, in both cases uniformly across time, thereby preventing κ_t from approaching unity. In the multiple-parameter setting, equation (17) essentially limits only the relative dynamics of $\{P_t\}$, preventing the penalization of different elements of the time-varying parameter from being too drastically different and varying too much across different time periods. Condition (17) is less stringent when the persistence in the prediction step is reduced (i.e., for Φ closer to $O_{K\times K}$) and/or when the strength of concavity is increased (i.e., for larger $\{\alpha_t\}$), as these conditions lead to stronger contractions (i.e., lower $\{\kappa_t\}$).

The presence of the scalar $\rho_t > 0$ in condition (17) indicates that the relative penalization between parameters matters, but not the overall magnitude. This is because a contraction in the norm $\|\cdot\|_P$ implies a contraction in the norm $\|\cdot\|_{\rho_t P}$ and vice versa. For this reason, the sufficient condition (17) is automatically satisfied if the sequence $\{P_t\}$ is a time-varying scalar multiple of a static matrix; i.e., $\{P_t = \zeta_t P\}$ for some sequence $\{\zeta_t > 0\}$ and $P \succ O_{K\times K}$ for which $P \succ \Phi' P \Phi$. Matrix A in condition (17) is included to ensure that the contraction coefficient with respect to the norm $\|\cdot\|_{\bar{P}}$ is bounded above, uniformly across time, at some value strictly below unity.

Result (18) implies the exponential almost sure (e.a.s.) convergence of the different paths $\{\theta_{t|t-1}\}$ and $\{\tilde{\theta}_{t|t-1}\}$ based on the same data, such that differences due to either (a) varying initializations $\theta_{0|0}$ and $\tilde{\theta}_{0|0}$ or (b) numerical errors due to finite computer precision disappear exponentially fast as time progresses. Importantly, Theorem 1 relies on the researcher-postulated, but not the true, observation density. Hence invertibility in the ProPar framework can be guaranteed without imposing additional restrictions on the true DGP, which is convenient as the true DGP is typically unknown. We may even allow Assumptions 1 to 5 to fail for a particular realization of the observation, as long as this violation occurs with probability zero. When the assumptions are guaranteed to hold for all observations y_t, the above stability result is entirely unaffected by model misspecification. In contrast, the contraction property of explicit score-driven models is typically contingent on the true DGP and the magnitude of the learning rate $H_t = P_t^{-1}$. The maximum-permitted learning rate in explicit score-driven models is closely tied to the properties of the true DGP, while an infringement of this (typically unknown) upper bound may yield an explosive recursion (e.g., Blasques et al., 2018, p. 1023).

In the optimization literature, similarly, Toulis and Airoldi (2017) find implicit stochastic-gradient algorithms to be convergent under arbitrary misspecification of the learning rate when the objective function is concave, whereas explicit methods require finetuning to avoid divergence.

3.2 Local information-theoretic optimality properties

To illustrate the optimality properties of our framework, we outline several desirable characteristics of the update procedure. We begin with investigating local optimality properties and subsequently investigate global behavior. First, Definition 1 introduces the concept of a likelihood-concordant update procedure.

Definition 1 (Likelihood concordance) *A parameter update from a prediction $\theta_{t|t-1}$ to an update $\theta_{t|t}$ based on the observation y_t is likelihood concordant if and only if $\log p(y_t|\theta_{t|t}) \geq \log p(y_t|\theta_{t|t-1})$. The update is strictly likelihood concordant if in addition $\log p(y_t|\theta_{t|t}) = \log p(y_t|\theta_{t|t-1})$ implies that $\theta_{t|t} = \theta_{t|t-1}$.*

In our view, likelihood concordance serves as a useful minimal requirement for a sensible parameter update. Specifically, if a parameter update is not likelihood concordant, the model fit evaluated at the observation y_t deteriorates after using y_t to generate the update, which is clearly undesirable. If Assumptions 1 and 2 hold, the ProPar update step is, due to the optimization (1), automatically strictly likelihood concordant. In contrast, while standard score-driven models with appropriately tuned learning rates may be likelihood concordant, the general class is not. This is because explicit-gradient methods cannot be guaranteed to improve the objective function unless the step size is arbitrarily small.

Likelihood concordance concerns an improvement in the likelihood of observing y_t, which is achieved by an updating scheme that utilizes the (same) observation y_t. The observation

y_t may be atypical, however, such that likelihood concordance, while desirable, does not necessarily imply an improvement in the expected likelihood of a theoretical redraw from the true density. Nevertheless, it turns out that we can guarantee an expected improvement in the likelihood for a new observation drawn from a set of positive probability in the vicinity of the observation y_t. To this end, we consider the Kullback-Leibler (KL) divergence of the predicted and updated densities, where both may be misspecified, relative to the true density. In computing the KL divergence, we consider only observations that are "similar" to y_t. Computing the difference between both KL divergences amounts to computing the difference in cross-entropies, relative to the true density, of the updated and predicted densities. Hence we define the local KL difference $\mathcal{D}_t(\mathcal{Y})$ with $\mathcal{Y} \subseteq \mathrm{Dom}(y) = \mathrm{Dom}(y_t)$ as

$$\mathcal{D}_t(\mathcal{Y}) \;:=\; \mathop{\mathbb{E}}_{y}\Big[\log p(y|\theta_{t|t}) - \log p(y|\theta_{t|t-1}) \,\Big|\, y \in \mathcal{Y}\Big], \tag{19}$$

where $\mathop{\mathbb{E}}_{y}[\cdot]$ denotes the expectation with respect to the true density $p_0(y|\theta_t^0)$. Here we distinguish between the actual y_t used to construct the update $\theta_{t|t}$, and a theoretical redraw from the true density, denoted y, which is assumed to be independent from y_t. We refer to updating schemes satisfying the condition $\mathcal{D}_t(\mathcal{Y}) > 0$, which is more stringent than likelihood concordance, as being locally KL-improving.

Definition 2 (Locally KL-improving updates) *A parameter update from prediction $\theta_{t|t-1}$ to update $\theta_{t|t}$ based on the observation y_t is locally KL-improving if and only if $\exists\delta \geq 0$ such that, for $\mathcal{Y} := \{y \in \mathrm{Dom}(y)|\ \|y - y_t\|^2 \leq \delta\ \}$, $\Pr(y \in \mathcal{Y}|\theta_t^0) := \int_{\mathcal{Y}} p_0(y|\theta_t^0)\mathrm{d}y > 0$ and $\mathcal{D}_t(\mathcal{Y}) > 0$.*

If the observations come from a discrete distribution and if $\theta_{t|t} \neq \theta_{t|t-1}$, then strict likelihood concordance trivially implies a local KL improvement. This is because, given that we have observed y_t, it is clear that $\Pr(y = y_t|\theta_t^0) > 0$. Hence we may pick $\delta = 0$ (which implies $\mathcal{Y} = y_t$) to obtain the desired result. If the observations take values in a continuum, we require the postulated density to be continuous, thus imposing no additional constraints on the DGP. For observations from a continuous distribution, under Assumptions 1 and 2 and requiring merely continuity of our postulated density, we can show that all non-trivial ProPar updates (i.e., for which $\theta_{t|t} \neq \theta_{t|t-1}$) represent local KL improvements.

Proposition 3 (Local KL improvement of the ProPar update) *Fix $t > 0$ and let Assumptions 1 and 2 hold. In addition, let either (a) $\Pr(y = y_t|\theta_t^0) > 0$ or (b) $p(y|\theta)$ be continuous in y, $\forall\theta \in \Theta$. Then, with probability one, the ProPar update from $\theta_{t|t-1}$ to $\theta_{t|t}$ using the observation y_t as in (1) is locally KL-improving if $\theta_{t|t} \neq \theta_{t|t-1}$.*

Our concept of a locally KL-improving update is related to that in Blasques et al. (2015), who introduce the notion of local realized KL optimality for univariate score-driven models with continuous observations. However, our setup is different in several ways. Our definition also encompasses discrete random variables and we limit neither y_t nor θ_t to the scalar case. The most important deviation is that we can dispense with the requirement in Blasques et al. (2015) that $\theta_{t|t}$ is contained in an arbitrarily small neighborhood of $\theta_{t|t-1}$. This condition, which effectively limits the approach to infinitesimally small step sizes, is unavoidable in explicit score-driven models because, more generally, explicit-gradient methods can only guarantee improvements of the objective function in the case of infinitesimal steps. In practice, the condition that the update $\theta_{t|t}$ remains arbitrarily close to the prediction $\theta_{t|t-1}$ requires that (a) the observation roughly confirms the accuracy of the prediction such that the update is only marginally different, or (b) a sizeable adjustment appears to be needed but the learning rate is kept arbitrarily small. These considerations suggest that information in explicit score-driven models may be slow to be incorporated; indeed, one of our empirical illustrations regarding the estimation of a time-varying market beta appears to confirm this (see Section 5.1).

In sum, we find that ProPar models possess a stronger and more generally applicable form of local optimality than explicit score-driven models. This result does not require us to place any additional demands on the likelihood or the DGP; indeed, Assumptions 1 and 2 and continuity of the postulated likelihood in the data are sufficient. The main disadvantage of explicit-gradient methods with non-infinitesimal learning rates, namely that they can lead to a deterioration of the objective function, is precluded when using implicit gradient methods. While it is tempting to try and prove a guaranteed global KL improvement of the update by investigating $\mathcal{D}_t(\mathcal{Y})$ with $\mathcal{Y} = \mathrm{Dom}(y_t)$, it is straightforward to show that this is generally infeasible due to the stochastic nature of the observation and, hence, the update.

3.3 Global optimality properties

While there is no hope of generalizing Proposition 3 to the global setting, here we demonstrate that the ProPar update is globally contracting towards some small region around the pseudo-true parameter. To this end, we make the following additional assumptions:

Assumption 6 (Uniqueness of pseudo-truth) *$\exists \theta_t^\star$ such that $\mathop{\mathbb{E}}_y[\log p(y|\theta_t^\star)] > \mathop{\mathbb{E}}_y[\log p(y|\theta)]$ $\forall \theta \in \Theta \setminus \{\theta_t^\star\}$ and $\mathop{\mathbb{E}}_y[\nabla(y|\theta_t^\star)] = 0$.*

Assumption 7 (Bounded information) *$\mathop{\mathbb{E}}_y[\|\nabla(y|\theta_t^\star)\|^2] < \infty$.*

Assumption 6 asserts the existence of a unique pseudo-truth $\theta_t^\star$ that maximizes the expected (postulated) log-likelihood function $\mathop{\mathbb{E}}_{y}[\log p(y|\theta_t^\star)]$. Equivalently, $\theta_t^\star$ is the unique minimizer of the KL divergence. If the logarithmic density is differentiable and strongly concave with probability one—i.e., Assumptions 4 and 5 hold for some $\alpha_t > 0$—then the existence of a unique pseudo-truth is automatic and need not be separately assumed. In the case of correct model specification, the truth and pseudo-truth coincide (i.e., $\theta_t^0 = \theta_t^\star$). Assumption 7 posits that the norm of the squared score computed with the postulated density, and evaluated in the pseudo-truth, is finite in expectation with respect to the true observation density.

When the prediction $\theta_{t|t-1}$ is very close to the pseudo-truth $\theta_t^\star$, the update $\theta_{t|t}$ will be inferior to the prediction with some positive probability, as $\theta_{t|t}$ is based on the noisy realization y_t. Hence an improvement is harder to achieve when the prediction is quite accurate; indeed, an improvement is impossible by Assumption 6 when the prediction is already pinpoint accurate (i.e., in the case $\theta_{t|t-1} = \theta_t^\star$). On the other hand, when the prediction $\theta_{t|t-1}$ is far from the pseudo-truth $\theta_t^\star$, the update $\theta_{t|t}$ will in expectation be superior to the prediction $\theta_{t|t-1}$. The next result makes explicit this tug of war between contractive and expansive forces.

Lemma 3 (Contractive and expansive forces) *Fix $t > 0$ and let Assumptions 1 to 7 hold. Then*

$$\begin{aligned}\underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\theta_{t|t}-\theta_t^\star\right\|_{P_t}^2\right]}_{\textit{MSE after updating}} \;\leq\; & \underbrace{\left\|\theta_{t|t-1}-\theta_t^\star\right\|_{P_t}^2}_{\textit{SE of prediction}} + \underbrace{2\mathop{\mathbb{E}}_{y_t}\left[\langle \nabla(y_t|\theta_{t|t})-\nabla(y_t|\theta_t^\star), \theta_{t|t}-\theta_t^\star\rangle\right]}_{\leq\, 0,\ \textit{contractive force}} \\ & + \underbrace{\mathop{\mathbb{E}}_{y_t}\left[\left\|\nabla(y_t|\theta_t^\star)\right\|_{H_t}^2\right]}_{\geq\, 0,\ \textit{expansive force}},\end{aligned} \tag{20}$$

where $\mathop{\mathbb{E}}_{y_t}[\cdot]$ denotes the expectation with respect to the true density $p_0(y_t|\theta_t^0)$ and (M)SE denotes the (mean) squared error.

Lemma 3 shows that the expected squared distance of the update from the pseudo-truth measured in a weighted norm (i.e., $\mathop{\mathbb{E}}_{y_t}[\|\theta_{t|t}-\theta_t^\star\|_{P_t}^2]$) is at most equal to the squared distance of the prediction from the pseudo-truth (i.e., $\|\theta_{t|t-1}-\theta_t^\star\|_{P_t}^2$) plus two additional terms. These terms determine whether the update is expected to be an improvement or not. When the researcher-postulated logarithmic density is concave with probability one (Assumption 5), we have that $\langle \nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_t^\star), \theta_{t|t}-\theta_t^\star\rangle$ is non-positive with probability one. Its expectation is then automatically non-positive, such that this term can be seen to act as a contractive force. The last term on the right-hand side involves a weighted norm of the postulated gradient evaluated at the pseudo-truth and averaged over y_t using the true density; hence, it reflects the irreducible noise obtained by updating based on the noisy observation y_t. Naturally, this term is non-negative and acts as an expansive force.

Importantly, the magnitude of the irreducible noise does not depend on the prediction $\theta_{t|t-1}$; hence, the strength of the expansive force remains constant as $\theta_{t|t-1}$ is moved further from the pseudo-truth $\theta_t^\star$. On the other hand, the contractive force is typically increasing in the distance of $\theta_{t|t-1}$ from $\theta_t^\star$, such that this contractive force tends to dominate when $\theta_{t|t-1}$ is far from $\theta_t^\star$. In the region where this contractive force dominates, we can expect updates to be beneficial. Conversely, the region around $\theta_t^\star$ where the expansive force dominates is known as the noise-dominated region (NDR, e.g. Ryu and Boyd, 2016, p. 15, Patrascu and Necoara, 2018, p. 3).

While the assumption of an increasing contractive force as we move further from the pseudo-truth is intuitive and verifiably true for most densities used in practice, it must still be formalized, as we do below in Assumption 8. The assumption itself is somewhat subtle to state, as it turns out that concavity (i.e., $\alpha_t = 0$) of the postulated logarithmic observation density is neither necessary nor sufficient, while strong concavity (i.e., $\alpha_t > 0$) is sufficient but stronger than necessary. Assumption 8 contains weaker versions of strong concavity similar to the ones employed in the optimization literature (e.g., Toulis et al., 2021, Assumption 3). Effectively, Assumption 8 ensures that the gradient is, on average, pointed in the correct direction while its magnitude increases sufficiently fast as we move away from the pseudo-truth.

Assumption 8 (Increasing expected gradient away from pseudo-truth)
a) $\exists \delta, C > 0$ *such that* $\forall \theta_{t|t} \in \Theta_\delta^\star := \{\theta \in \Theta | \, \|\theta_t^\star - \theta\|^2 \geq \delta\}$,

$$2\underset{y_t}{\mathbb{E}}\Big[\langle \nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_t^\star), \theta_{t|t} - \theta_t^\star\rangle\Big] \; < \; -\underset{y_t}{\mathbb{E}}\Big[\|\nabla(y_t|\theta_t^\star)\|_{H_t}^2\Big] - C, \tag{21}$$

b) $\exists \tilde{\alpha}_t > 0$ *such that* $\forall \theta_{t|t} \in \Theta$,

$$\underset{y_t}{\mathbb{E}}\Big[\langle \nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_t^\star), \theta_{t|t} - \theta_t^\star\rangle\Big] \; \leq \; -\tilde{\alpha}_t \underset{y_t}{\mathbb{E}}\Big[\|\theta_{t|t} - \theta_t^\star\|^2\Big]. \tag{22}$$

Assumption 8a posits that if $\theta_{t|t}$ is far enough from $\theta_t^\star$, the contractive force dominates the irreducible noise $\underset{y_t}{\mathbb{E}}[\|\nabla(y_t|\theta_t^\star)\|_{H_t}^2]$ by at least some positive amount C. Assumption 8b is a stronger version of Assumption 8a and assumes that the contractive force scales with the distance $\|\theta_t^\star - \theta_{t|t}\|^2$. Assumption 8b can in turn be seen as a weaker condition than α_t-strong concavity, as the latter implies the existence of some $\tilde{\alpha}_t \geq \alpha_t$. This is because the relationship in Assumption 8b is expressed (a) in terms of an expectation and (b) in relation only to the pseudo-truth $\theta_t^\star$, whereas α_t-strong concavity would require a similar inequality to hold (a) with probability one and (b) for all pairs of points. Assumptions 8a and 8b thus allow for some degree of non-concavity in the postulated log-likelihood function; combining them with Lemma 3 yields the contraction result in Theorem 2.

Theorem 2 (Contraction to the NDR) *Fix $t > 0$ and let Assumptions 1-4, 6-7 and 8a hold. Then $\exists \delta, C > 0$ such that $\forall \theta_{t|t} \in \Theta^\star_\delta := \{\theta \in \Theta | \, \|\theta^\star_t - \theta\|^2 \geq \delta\}$,*

$$\mathop{\mathbb{E}}_{y_t}\Big[\|\theta_{t|t} - \theta^\star_t\|^2_{P_t}\Big] \;\leq\; \|\theta_{t|t-1} - \theta^\star_t\|^2_{P_t} - C. \tag{23}$$

If in addition 8b holds for some $\tilde{\alpha}_t > 0$, then $\forall \theta_{t|t} \in \Theta$,

$$\mathop{\mathbb{E}}_{y_t}\Big[\|\theta_{t|t} - \theta^\star_t\|^2_{P_t + 2\tilde{\alpha}_t I_K}\Big] \;\leq\; \|\theta_{t|t-1} - \theta^\star_t\|^2_{P_t} + \mathop{\mathbb{E}}_{y_t}\Big[\|\nabla(y_t|\theta^\star_t)\|^2_{H_t}\Big]. \tag{24}$$

In Theorem 2, Assumption 8a guarantees a fixed reduction in the expected squared error when the prediction is far from the pseudo-truth. Under Assumption 8b, this result can be strengthened to obtain a global linear contraction up to some level of accuracy determined by the weighted magnitude of the additive noise. The speed of contraction is regulated by the average curvature of the log-likelihood function from the prediction to the pseudo-truth, measured by $\tilde{\alpha}_t$, and the size of the penalty P_t. Ceteris paribus, a smaller penalty or stronger form of concavity yields a faster contraction. Moreover, the irreducible noise is increasing in the size of the learning-rate matrix $H_t = P_t^{-1}$, such that larger learning rates (or, equivalently, smaller penalties) lead to a larger NDR. The optimal choice of learning rate is therefore determined by a trade-off between contraction speed when far from the pseudo-truth and the size of the NDR. By continuity, the expected contraction in terms of the parameter θ to the pseudo-truth $\theta^\star_t$ also implies a contraction on an upper bound in the expected log-likelihood difference relative to the pseudo-truth; in the correctly specified case, this is the KL divergence. We conclude that the ProPar update possesses advantageous optimality properties that are unavailable in the explicit domain; this is the strength of preserving and fully exploiting all information in the log-likelihood contribution using optimization.

4 Estimation

The parameters of the ProPar model, including the penalty matrices $\{P_t\}$ in the update (1), parameters ω and Φ in the prediction step (4), and any additional fixed shape parameters in the observation density are generally unknown and need to be estimated. In our empirical illustrations below, the penalty matrix is taken to be constant (i.e., $P_t = P$ for all t) and targeting is used for the initialization (alternatively, the initial parameter values $\theta_{0|0}$ could have been estimated). Determination of all aforementioned parameters can proceed by maximum-likelihood (ML) estimation based on the standard prediction-error decomposition. We use the results obtained in Blasques et al. (2022), who derive sufficient conditions for consistency and asymptotic normality of the ML estimator for explicit score-driven models with a scalar time-varying parameter ($K = 1$). They consider both the correctly and incorrectly specified cases. A crucial ingredient of their proofs is the invertibility concept in Bougerol (1993) and Straumann and Mikosch (2006).

In the asymptotic ML theory of Blasques et al. (2022), verifying the contraction condition required for filter invertibility is often the hardest part. For concave logarithmic observation densities, Theorem 1 presents a simpler and stronger form of invertibility for the ProPar model than is available for explicit score-driven models. Therefore, under similar or possibly weaker assumptions regarding the DGP and the parameter space of P_t and Φ, we may obtain consistency and asymptotic normality of the ML estimator by applying the theory developed in Blasques et al. (2022). Specifically, these assumptions include that the static parameters to be estimated are identified, that the series $\{y_t\}$ is stationary ergodic and near-epoch dependent with some finite moments, and that the postulated density is sufficiently continuous in its arguments and has bounded derivatives. The latter conditions provide sufficient moments to be used in the appropriate law of large numbers and central-limit theorem (see Blasques et al., 2022, Theorem 4.6 and 4.15 for details). For log-concave densities, we conjecture that for ProPar models these results can in principle be straightforwardly extended to the multi-parameter case ($K > 1$), a full asymptotic investigation of which is beyond the scope of this article.

5 Empirical illustrations

5.1 Linear regression with time-varying slope

The capital asset pricing Model (CAPM), an important benchmark in finance, links the expected excess returns of individual assets to those of the market in a linear fashion. However, empirical evidence (e.g., Jagannathan and Wang, 1996) shows that the assumption of a constant market coefficient β may be unrealistic, especially in equity markets. We examine the possible time-varying nature of the CAPM market β using the ProPar framework. We model the excess asset return y_t as

$$y_t \;=\; \alpha \,+\, \beta_t\, m_t \,+\, \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \tag{25}$$

where α is a static intercept, m_t denotes the excess market return at time t, and ε_t is an i.i.d. normally distributed shock with mean zero and variance σ^2. The ProPar update is a special case of the general setup of Example 1. For the prediction step, we use the linear first-order specification (4). The penalty parameter or its inverse, the learning rate $\eta > 0$, is assumed to be constant.

We apply the ProPar dynamic regression model (25) to simple daily excess returns of Microsoft (MSFT) from 14 March 1986 until 29 April 2022, obtained from Yahoo Finance.[1] For the market return and risk-free rate we use the series from Kenneth French's database.[2] Figure 1 shows the evolution of $\beta_{t|t-1}$ for the ProPar model and its explicit version, i.e., the explicit score-driven model. Figure 1 also contains the estimated impact curves $\beta_{t|t} - \beta_{t|t-1}$ with respect to the market return m_t for a fixed $y_t = 0$ and two different predictions (i.e., $\beta_{t|t-1} = 1$ and $\beta_{t|t-1} = -0.5$).

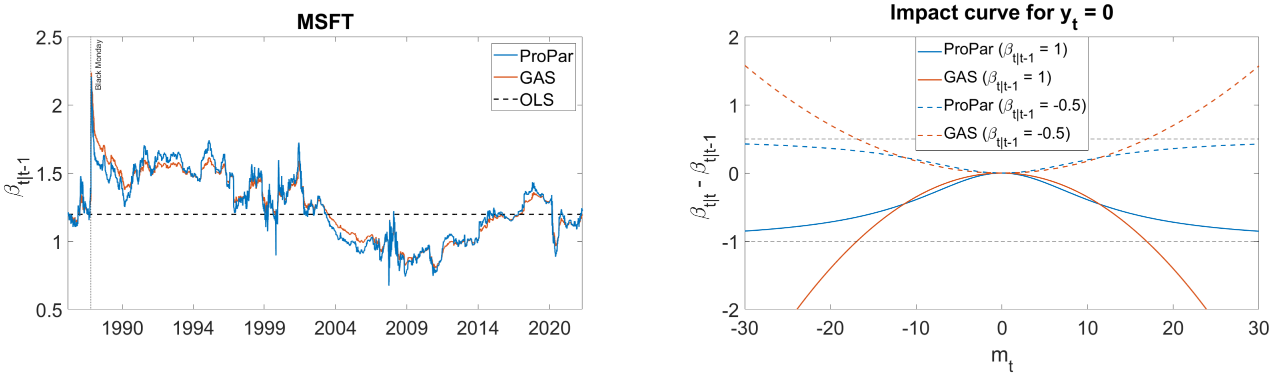

Figure 1: Time-evolution of $\beta_{t|t-1}$ and estimated impact curve of the ProPar and explicit score-driven (i.e., GAS) models for MSFT from March 1986 until April 2022. Vertical dotted lines mark Black Monday on October 19, 1987.

In Figure 1, we observe that the ProPar and explicit score-driven models generally generate a similar series $\{\beta_{t|t-1}\}$, while the path generated by the ProPar model seems to be leading. In particular, the explicit score-driven model appears to be slow to recover from large shocks, such as the crash on Black Monday, 1987. The reason for this delayed reaction is that, in explicit score-driven models, the learning rate η must be substantially reduced to deal with outliers, even though this implies a reduced responsiveness in the remainder of the sample, as is evident around 1994 and 2004. This problem is drastically reduced for the ProPar model by the more favorable (asymptotic) impact curve with respect to the exogenous input. In Figure 1, we observe an unbounded quadratic impact of m_t on the adjustment $\beta_{t|t} - \beta_{t|t-1}$ in the explicit score-driven model, while the ProPar impact curve is similar for

[1] https://finance.yahoo.com/quote/MSFT/history?p=MSFT

[2] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

small $|m_t|$ but bounded for large $|m_t|$. Specifically, for ProPar we observe that $|m_t| \to \infty$ implies $\beta_{t|t} - \beta_{t|t-1} \to -\beta_{t|t-1}$ and hence $\beta_{t|t} \to 0$. When the exogenous variable is excessively large, therefore, the dynamic slope $\beta_{t|t}$ under the ProPar specification reverts to zero. This enhanced stability property allows ProPar's estimated learning rate to substantially exceed that of the explicit score-driven model ($\hat{\eta} = 0.0169$ versus $\hat{\eta} = 0.0092$ for ProPar and the explicit method, respectively), which explains ProPar's higher sensitivity during non-crisis times.

5.2 Time-varying volatility

Modeling asset-price volatility plays a central role in finance and provides important input for risk management, among others. We consider a time-varying volatility model using the ProPar filter. Specifically, we model the logarithmic asset return y_t as

$$y_t = \mu + \sigma_t \, z_t, \qquad \sigma_t = \exp(h_t) \qquad z_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1), \tag{26}$$

where μ is a static mean, $h_t = \log \sigma_t$ is the dynamic conditional logarithmic volatility, and z_t is a standardized i.i.d. normally distributed shock. The ProPar update and prediction are

$$h_{t|t} = h_{t|t-1} + \eta \left[\left(\frac{y_t - \mu}{\exp(h_{t|t})} \right)^2 - 1 \right], \qquad h_{t+1|t} = \omega + \phi \, h_{t|t}, \tag{27}$$

where the prediction parameters (i.e., $\omega \in \mathbb{R}$ and $\phi \in [0,1)$) and the learning rate (i.e., $\eta > 0$) are to be estimated by maximum likelihood. The ProPar update $h_{t|t}$ can be analytically solved from equation (27) using the Lambert W function, which is available in most standard software packages.

We estimate the ProPar volatility model in (27) for daily S&P500 returns from 4 January 2000 until 28 June 2022, obtained from the Oxford-Man library. We compare the ProPar model against its explicit (i.e., GAS) counterpart, which can be obtained by replacing $h_{t|t}$ on the right-hand-side of update (27) by $h_{t|t-1}$. In addition, we estimate Nelson's (1991) EGARCH model without a leverage term. Figure 2 shows the estimated paths of $\sigma_{t|t-1}$ for the different models and the estimated impact curves for $\sigma_{t|t-1} = 1$.

Figure 2 reveals that the ProPar filter closely aligns with the EGARCH model. The explicit score-driven model displays a similar pattern, but is more sensitive to large shocks during low-volatility periods. The estimated impact curves demonstrate that the ProPar model is less sensitive to large shocks than the EGARCH model, which in turn is more robust than the explicit score-driven model. In the absence of large shocks, the three models closely

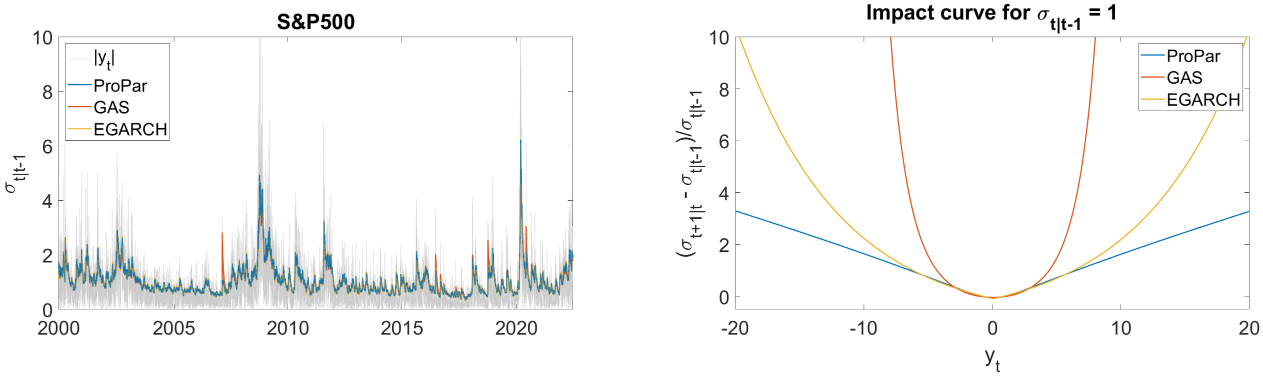

Figure 2: Time-evolution of $\sigma_{t|t-1}$ and estimated impact curve of the ProPar, explicit score-driven (i.e., GAS) and EGARCH models for daily S&P500 returns from January 2000 until June 2022.

align. In terms of fit, we find minor advantages of the ProPar model over the EGARCH model, which in turn outperforms the explicit score-driven model. For example, the log-likelihood values are -7781.4, -7790.3 and -7795.7 for the ProPar, EGARCH, and explicit score-driven models, respectively. We find a similar result in terms of the mean squared error (MSE) when compared to the 5-minute realized variance (5.033, 5.047, and 5.120, respectively). Similarly, the ProPar update $\sigma^2_{t|t}$ outperforms its explicit counterpart (MSE 4.902 versus 4.972, respectively), which suggests that the ProPar updating step may be useful for now-casting. We conclude that the ProPar framework can be used to construct a competitive volatility model with an ingrained robustness to outliers, even when it is based on a Gaussian observation density. This illustrates that the robustness of ProPar comes about by a different mechanism than in explicit score-driven models, which typically require heavy-tailed observation densities.

5.3 Time-varying growth at risk

Modeling macroeconomic downside risk is crucial for policymakers. The growth-at-risk (GaR) framework refers to conditional lower quantiles of GDP growth and has become a popular measure for macroeconomic risk assessment. Typically, estimation is performed by means of quantile regressions (QRs; see Koenker and Hallock, 2001). These regressions usually rely on a set of exogenous variables; e.g., capturing the relationship between GaR on the one hand and economic and financial conditions on the other (Adrian et al., 2019).

We propose to endogenously update a time-varying conditional quantile by postulating an asymmetric Laplace distribution with a time-varying location. Maximizing such a density is equivalent to the minimization of Koenker and Bassett's (1978) QR check function, see Koenker and Machado, 1999. The ProPar update for the τ-level quantile at time t, denoted by $q_{t|t}(\tau)$, can be computed in closed form as

$$q_{t|t}(\tau) = 1[y_t \leq q_{t|t-1}(\tau)]\ \max\{y_t, q^{\mathrm{e}}_{t|t}(\tau)\} + 1[y_t > q_{t|t-1}(\tau)]\ \min\{y_t, q^{\mathrm{e}}_{t|t}(\tau)\}, \tag{28}$$

where y_t denotes the GDP growth rate at time t, while $1[\cdot]$ equals an indicator function that equals one if the condition in square brackets is satisfied and zero otherwise. The ProPar update (28) is expressed in terms of the explicit score-driven update, denoted $q^{\mathrm{e}}_{t|t}(\tau)$, which is obtained as follows:

$$q^{\mathrm{e}}_{t|t}(\tau) = q_{t|t-1}(\tau) + \frac{\eta}{\sigma}(\tau - 1[y_t \leq q_{t|t-1}(\tau)]), \tag{29}$$

where $\eta > 0$ and $\sigma > 0$ denote the learning rate and a dispersion parameter, respectively, which are assumed constant over time. The form of $q^{\mathrm{e}}_{t|t}(\tau)$ is the same as in Engle and Manganelli's (2004) adaptive CAViaR model, yielding a downward adjustment of size $\eta(\tau - 1)/\sigma$ when the observed growth y_t falls below the quantile prediction $q_{t|t-1}$ and an upward adjustment of size $\eta\tau/\sigma$ otherwise. Equation (28) reveals that the ProPar update is a shrunken version of the explicit update; in particular, ProPar has the desirable property that the update can never be more extreme than (i.e., is capped at) the observation y_t.

Quantile crossing poses an important problem in practice when simultaneously modeling multiple quantiles using QRs. Thanks to the particular form of the update (28), the ProPar model can ensure an appropriate ordering of the quantiles using simple parameter restrictions. Specifically, if we assume that all quantile updates share the same learning rate η and dispersion parameter σ, then the updated quantiles remain correctly ordered. To illustrate, consider an observation y_t that falls between the predictions of two different quantiles. Consequently, one must be updated downward, the other upward. Because the ProPar update is capped at the observation, the two quantiles cannot cross. In contrast, the explicit score-driven update generally permits such crossings to occur. To guarantee that the correct ordering of quantiles is maintained not only in the update but also in the prediction step, we specify the prediction as

$$q_{t+1|t}(\tau) \;=\; c(\tau)\,(1-\phi) \;+\; \phi\, q_{t|t}(\tau) \;+\; \gamma\, x_t, \tag{30}$$

with an autoregressive parameter $\phi \in [0, 1)$ that is common across quantiles and intercepts $c(\tau)$ that are strictly ordered in τ. Furthermore, x_t denotes an exogenous variable available at time t with common slope parameter γ. While it may be useful to allow for different sensitivities to the exogenous input x_t for different quantiles, this has the potential to introduce quantile crossings. Moreover, we find for our application that the likelihood improvement of quantile specific slopes $\gamma(\tau)$ is too small to justify the additional model complexity.

We estimate the 5, 10, 25, and 50 percent GaR using the ProPar and adaptive CAViaR models using quarterly US GDP growth rates from 1971-Q1 until 2021-Q4. For the exogenous variable x_t, we follow Adrian et al. (2019) in using the National Financial Conditions Index (NFCI), where quarterly values are constructed by averaging the corresponding weekly values. Both time series were obtained from the FRED database.[3] To reduce the number of parameters to be estimated, we use a targeting approach and set $c(\tau)$ equal to the corresponding empirical quantiles. The remaining static parameters are estimated in a composite-likelihood fashion, comparable to Zou and Yuan (2008). We fix the scale parameter $\sigma = 1$, as it does not influence the quantile dynamics and can be treated as a nuisance parameter (e.g., Geraci and Bottai, 2007). Our postulated log-likelihood function equals the sum of four logarithmic Laplace densities, of which three are asymmetric and one is symmetric (i.e., the one corresponding to the median).

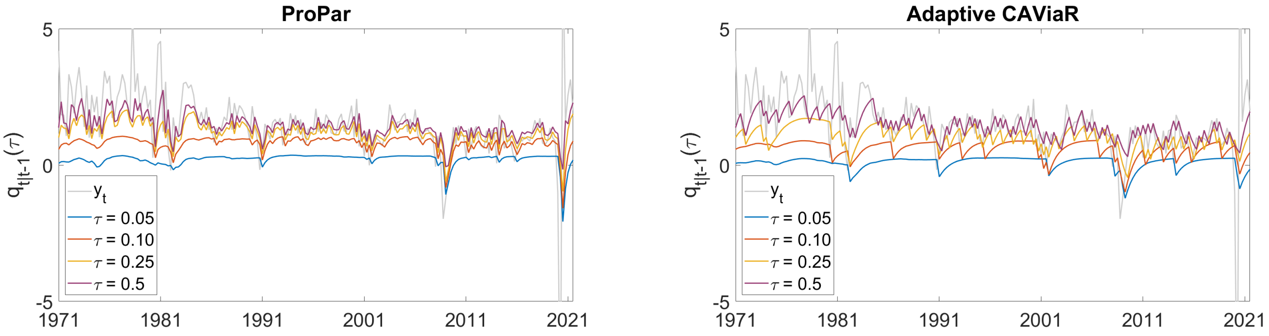

Figure 3: Growth-at-risk estimates for the ProPar and adaptive CAViaR models for $\tau = 0.05$, $\tau = 0.10$, $\tau = 0.25$ and $\tau = 0.50$, 1971-Q1 until 2021-Q4.

Figure 3 shows the 5, 10, 25, and 50 percent GaR estimates obtained from the ProPar model (28) and adaptive CAViaR model (29). It reveals that the ProPar model is more responsive than the adaptive CAViaR model. For example, the ProPar model shows greater downward adjustments during the onset of the COVID-19 pandemic in April 2020 in combination with a faster mean reversion after the crisis. This behavior is made possible by the enhanced stability of the implicit update (28) relative to the explicit update (29), which means that the estimated learning rate η of the ProPar model much exceeds that of the adaptive CAViaR model ($\hat{\eta} = 4.002$ and $\hat{\eta} = 0.804$ for ProPar and adaptive CAViaR, respectively). Furthermore, the adaptive CAViaR model occasionally suffers from quantile crossing, while all ProPar quantiles remain strictly ordered at all times. When considering the median ($\tau = 0.50$), the adaptive CARiaR update frequently cuts across (i.e., overshoots)

[3] See https://fred.stlouisfed.org/series/GDP and https://fred.stlouisfed.org/series/NFCI.

the observation y_t, whereas the ProPar model, with its capped updates, produces more stable dynamics that closely mimic the data. In line with Adrian et al.'s (2019) results, we find that the effect of the NFCI on the quantiles is negative ($\hat{\gamma} = -0.052$ and $\hat{\gamma} = -0.019$ for ProPar and adaptive CAViaR, respectively), such that higher values of the NFCI correspond to more negative quantiles.

6 Conclusion

This article introduced a novel framework for updating time-varying parameters in an observation-driven setting. Specifically, we proposed a proximal-parameter update that maximizes, at each point in time, the logarithmic observation density subject to a quadratic penalty centered at the one-step-ahead prediction. The first-order condition associated with this maximization can be written as an implicit stochastic-gradient update, connecting the proposed method with recent advances in statistics and machine learning. We derived model invertibility for the class of (possibly misspecified) concave logarithmic observation densities and formulated sufficient conditions for a global contraction of the parameter update towards a pseudo-truth. We demonstrated that the class of explicit score-driven models—known variously as dynamic conditional score (DCS; Harvey, 2013) or generalized autoregressive score (GAS; Creal et al., 2013) models—can be obtained within the ProPar framework by replacing the logarithmic observation density at each point in time by its local-linear approximation around the prediction. More directly, this class of models can be obtained by replacing ProPar's implicit stochastic-gradient update with its explicit version. Comparing the two methods, we found that the ProPar model extends several attractive properties of explicit score-driven models from the local to the global setting. In addition, it admits stronger contraction properties, yielding a well-behaved filter regardless of multiple types of misspecification. Empirical benefits were demonstrated in three illustrations involving asset pricing, stock-market volatility, and growth-at-risk.

References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). Vulnerable growth. *American Economic Review* **109**, 1263–89.

Anderson, Brian DO and John B Moore (2012). Optimal filtering. Prentice-Hall.

Asi, Hilal and John C Duchi (2019). Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization* **29**, 2257–2290.

Bianchi, Pascal (2016). Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization* **26**, 2235–2260.

Blasques, Francisco, Janneke van Brummelen, Siem Jan Koopman, and Andre Lucas (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics* **227**, 325–346.

Blasques, Francisco, Paolo Gorgi, Siem Jan Koopman, and Olivier Wintenberger (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics* **12**, 1019–1052.

Blasques, Francisco, Siem Jan Koopman, and Andre Lucas (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* **102**, 325–343.

Bougerol, Philippe (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization* **31**, 942–959.

Cox, David R, Gudmundur Gudmundsson, Georg Lindgren, Lennart Bondesson, Erik Harsaae, Petter Laake, Katarina Juselius, and Steffen L Lauritzen (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**, 777–795.

Creal, Drew, Bernd Schwaab, Siem Jan Koopman, and Andre Lucas (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics* **96**, 898–915.

Engle, Robert F and Simone Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* **22**, 367–381.

Geraci, Marco and Matteo Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**, 140–154.

Gorgi, Paolo (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1325–1347.

Grimmer, Benjamin, Haihao Lu, Pratik Worah, and Vahab Mirrokni (2022). The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, Advance online publication.

Hare, Warren and Claudia Sagastizábal (2009). Computing proximal points of nonconvex functions. *Mathematical Programming* **116**, 221–258.

Harvey, Andrew (2013). Dynamic models for volatility and heavy tails: With applications to financial and economic time series. Vol. **52**. Cambridge University Press.

Harvey, Andrew and Rutger-Jan Lange (2017). Volatility modeling with a generalized t distribution. *Journal of Time Series Analysis* **38**, 175–190.

Harvey, Andrew and Alessandra Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association* **109**, 1112–1122.

Jagannathan, Ravi and Zhenyu Wang (1996). The conditional CAPM and the cross-section of expected returns. *The Journal of Finance* **51**, 3–53.

Kalman, Rudolph Emil (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

Koenker, Roger and Gilbert Bassett (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* **46**, 33–50.

Koenker, Roger and Kevin F Hallock (2001). Quantile regression. *Journal of Economic Perspectives* **15**, 143–156.

Koenker, Roger and Jose AF Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**, 1296–1310.

Koopman, Siem Jan, Andre Lucas, and Marcel Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics* **98**, 97–110.

Kulis, Brian and Peter L Bartlett (2010). Implicit online learning. *Proceedings of the 27th International Conference on Machine Learning*, 575–582.

Li, Mu, Tong Zhang, Yuqiang Chen, and Alexander J Smola (2014). Efficient mini-batch training for stochastic optimization. *Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining*, 661–670.

Nelson, Daniel B (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 347–370.

Opschoor, Anne, Pawel Janus, André Lucas, and Dick Van Dijk (2018). New HEAVY models for fat-tailed realized covariances and returns. *Journal of Business & Economic Statistics* **36**, 643–657.

Patrascu, Andrei and Ion Necoara (2018). Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *The Journal of Machine Learning Research* **18**, 7204–7245.

Rockafellar, R Tyrrell (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14**, 877–898.

Ryu, Ernest K and Stephen Boyd (2016). Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. *Author website: https://web.stanford.edu/boyd/papers/pdf/spi.pdf.*

Stock, James H and Mark W Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics* **14**, 11–30.

Straumann, Daniel and Thomas Mikosch (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics* **34**, 2449–2495.

Teräsvirta, Timo (2009). An introduction to univariate GARCH models. *Handbook of Financial Time Series*. Springer, 17–42.

Toulis, Panos and Edoardo M Airoldi (2015). Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Statistics and Computing* **25**, 781–795.

— (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics* **45**, 1694–1727.

Toulis, Panos, Thibaut Horel, and Edoardo M Airoldi (2021). The proximal Robbins-Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**, 188–212.

Toulis, Panos, Dustin Tran, and Edoardo M Airoldi (2016). Towards stability and optimality in stochastic gradient descent. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* **51**, 1290–1298.

Zou, Hui and Ming Yuan (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108–1126.

Appendix to Robust Observation-Driven Models Using Proximal-Parameter Updates

A Example 1: Linear regression

Consider the linear regression model with dependent variable $y_t \in \mathbb{R}$ and independent variable $x_t \in \mathbb{R}^K$, that is,

$$y_t \;=\; \beta_t' x_t \,+\, \varepsilon_t, \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \tag{A.1}$$

where β_t is a $K \times 1$ vector of time-varying parameters and ε_t is an i.i.d. normally distributed innovation with variance σ^2.

The log-likelihood contribution $\log p(y_t|\beta)$ is obviously twice continuously differentiable with respect to β for all y_t, such that Assumption 4 (differentiability) holds. In addition, the Hessian is equal to $-\frac{1}{\sigma^2} x_t x_t'$ and is therefore negative semi-definite. Combined with strong concavity of the penalty this means that the regularized log likelihood $f(\beta|y_t, \beta_{t|t-1}, P_t) := \log p(y_t|\beta) - \frac{1}{2}\left\|\beta - \beta_{t|t-1}\right\|^2_{P_t}$ is strongly concave in β. Because $f(\beta|y_t, \beta_{t|t-1}, P_t)$ is finite-valued for any $\beta \in \mathbb{R}^K$, we have that it is thus strictly proper concave such that Assumption 2 (strictly concave regularized log likelihood) holds.

The first-order condition (FOC) of the ProPar update at time t associated with the model (A.1) takes the following form

$$\beta_{t|t} = \beta_{t|t-1} + H_t \nabla(y_t|\beta_{t|t}, x_t), \tag{A.2}$$

where $H_t = P_t^{-1}$ is the learning-rate matrix and $\nabla(y_t|\beta_{t|t}, x_t)$ denotes the implicit score given as,

$$\nabla(y_t|\beta_{t|t}, x_t) \;=\; \frac{y_t - \beta_{t|t}' x_t}{\sigma^2}\, x_t. \tag{A.3}$$

Note that strong concavity of $f(\beta|y_t, \beta_{t|t-1}, P_t)$ and the unrestricted nature of the optimization (i.e. we maximize over $\mathbb{R}^K$) imply that if the FOC (A.2) has a solution then it is the unique global maximizer. Solving the FOC will thus also directly verify Assumptions 1 (existence) and 3 (interior solution).

Collecting all terms containing $\beta_{t|t}$ on the left-hand side, we may write the FOC in (A.2) as

$$(I_K + H_t \frac{x_t x_t'}{\sigma^2})\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2}. \tag{A.4}$$

Now using the Sherman-Morrison identity, we left-multiply with $(I_K + H_t \frac{x_t x_t'}{\sigma^2})^{-1} = I_K - \frac{H_t x_t x_t'}{\sigma^2 + x_t' H_t x_t}$, which yields

$$\beta_{t|t} = (I_K - \frac{H_t x_t x_t'}{\sigma^2 + x_t' H_t x_t})(\beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2}). \tag{A.5}$$

Eliminating brackets and using the notation $\|x_t\|^2_{H_t} := x_t' H_t x_t$ then gives

$$\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t x_t}{\sigma^2} - \frac{H_t x_t x_t'}{\sigma^2 + \|x_t\|^2_{H_t}} \beta_{t|t-1} - \frac{H_t x_t x_t'}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t x_t}{\sigma^2}, \tag{A.6}$$

where changing the ordering using the fact that y_t, σ^2, $x_t'\beta_{t|t-1}$ and $\|x_t\|^2_{H_t}$ are scalars and again using the definition of $\|x_t\|^2_{H_t}$, we get

$$\beta_{t|t} = \beta_{t|t-1} + H_t \frac{y_t}{\sigma^2} x_t - \frac{1}{\sigma^2 + \|x_t\|^2_{H_t}} H_t x_t' \beta_{t|t-1} x_t - \frac{\|x_t\|^2_{H_t}}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t}{\sigma^2} x_t. \tag{A.7}$$

Multiplying the second and third term on the right-hand side with $\frac{\sigma^2 + \|x_t\|^2_{H_t}}{\sigma^2 + \|x_t\|^2_{H_t}}$ and $\frac{\sigma^2}{\sigma^2}$, respectively, allows us to combine the second through fourth terms as follows

$$\beta_{t|t} = \beta_{t|t-1} + \frac{\sigma^2}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \frac{y_t - x_t'\beta_{t|t-1}}{\sigma^2} x_t, \tag{A.8}$$

where using the definition of the explicit gradient $\nabla(y_t|\beta_{t|t-1}, x_t)$ gives the final result

$$\beta_{t|t} = \beta_{t|t-1} + \frac{\sigma^2}{\sigma^2 + \|x_t\|^2_{H_t}} H_t \nabla(y_t|\beta_{t|t-1}, x_t). \tag{A.9}$$

B Proofs

B.1 Proposition 1: Gradient alignment

By Assumption 2 we have that the regularized log likelihood $f(\theta|y_t, \theta_{t|t-1})$ is concave in θ with probability one in y_t. As a result, we have for almost every y_t that

$$f(\theta_{t|t}|y_t, \theta_{t|t-1}) \leq f(\theta_{t|t-1}|y_t, \theta_{t|t-1}) + \langle \nabla(y_t|\theta_{t|t-1}), \theta_{t|t} - \theta_{t|t-1} \rangle, \tag{B.10}$$

reordering and using the fact that $\theta_{t|t}$ maximizes $f(\theta|y_t, \theta_{t|t-1})$ we obtain

$$\langle \nabla(y_t|\theta_{t|t-1}), \theta_{t|t} - \theta_{t|t-1} \rangle \geq f(\theta_{t|t}|y_t, \theta_{t|t-1}) - f(\theta_{t|t-1}|y_t, \theta_{t|t-1}) \geq 0, \tag{B.11}$$

which yields the desired result. Filling in the first-order condition produces

$$\langle \nabla(y_t|\theta_{t|t-1}), H_t \nabla(y_t|\theta_{t|t}) \rangle \geq 0, \tag{B.12}$$

providing an equivalent statement under Assumptions 3 and 4.

B.2 Corollary 1: Gradient-sign concordance in one dimension

Using the result of Proposition 1, we have in the scalar case that $\nabla(y_t|\theta_{t|t-1})\nabla(y_t|\theta_{t|t}) \geq 0$ using the strict positivity of the learning rate. Furthermore, $\nabla(y_t|\theta_{t|t}) = 0$ implies that $\theta_{t|t} = \theta_{t|t-1}$ by the first-order condition, in turn implying that $\nabla(y_t|\theta_{t|t-1}) = \nabla(y_t|\theta_{t|t}) = 0$. Conversely, if $\nabla(y_t|\theta_{t|t-1}) = 0$, we have that $\theta_{t|t} = \theta_{t|t-1}$, as filling in $\theta_{t|t-1}$ solves the first-order condition (and Assumption 2 implies uniqueness of $\theta_{t|t}$). Therefore, $\nabla(y_t|\theta_{t|t-1}) = 0$ if and only if $\nabla(y_t|\theta_{t|t}) = 0$. Combining this with the fact that $\nabla(y_t|\theta_{t|t-1})\nabla(y_t|\theta_{t|t}) \geq 0$, we obtain $\text{sgn}(\nabla(y_t|\theta_{t|t})) = \text{sgn}(\nabla(y_t|\theta_{t|t-1}))$.

B.3 Proposition 2: Step-size shrinkage

Using the first-order conditions of the implicit and explicit update we obtain that the difference in the update $\theta_{t|t} - \theta^{\text{e}}_{t|t}$ can be written as

$$\theta_{t|t} - \theta^{\text{e}}_{t|t} = \theta_{t|t-1} + H_t\nabla(y_t|\theta_{t|t}) - \theta_{t|t-1} - H_t\nabla(y_t|\theta_{t|t-1}), \tag{B.13}$$

whereby rearranging yields

$$\theta^{\text{e}}_{t|t} - \theta_{t|t-1} = \theta_{t|t} - \theta_{t|t-1} - H_t[\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1})]. \tag{B.14}$$

Pre-multiplying with $H_t^{-1/2} = P_t^{1/2}$, which denotes the symmetric square root of $H_t^{-1} = P_t$, and taking the quadratic norm yields

$$\begin{aligned} \|\theta^{\text{e}}_{t|t} - \theta_{t|t-1}\|^2_{P_t} = \|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t} &- 2\langle\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1}), \theta_{t|t} - \theta_{t|t-1}\rangle \\ &+ \|\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1})\|^2_{H_t}. \end{aligned} \tag{B.15}$$

Now using that $\|\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1})\|^2_{H_t} \geq 0$ and that $\langle\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_{t|t-1}), \theta_{t|t} - \theta_{t|t-1}\rangle \leq -\alpha_t\|\theta_{t|t} - \theta_{t|t-1}\|^2$ by (strong) concavity of the log likelihood from Assumption 5, we obtain

$$\|\theta_{t|t} - \theta_{t|t-1}\|^2_{P_t+2\alpha_t I_K} \leq \|\theta^{\text{e}}_{t|t} - \theta_{t|t-1}\|^2_{P_t}, \tag{B.16}$$

which concludes the proof.

B.4 Lemma 1: Prediction-to-update stability

Consider two predictions $\theta_{t|t-1}$ and $\tilde{\theta}_{t|t-1}$ that are updated based on the observation y_t to $\theta_{t|t}$ and $\tilde{\theta}_{t|t}$, respectively. We consider the weighted norm with respect to P_t of the difference in updates and substitute the first-order conditions. This yields

$$\begin{aligned}\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} &= \langle P_t\tilde{\theta}_{t|t-1} + \nabla(y_t|\tilde{\theta}_{t|t}) - P_t\theta_{t|t-1} - \nabla(y_t|\theta_{t|t}), \tilde{\theta}_{t|t} - \theta_{t|t}\rangle \\ &= \langle P_t(\tilde{\theta}_{t|t-1} - \theta_{t|t-1}), \tilde{\theta}_{t|t} - \theta_{t|t}\rangle + \langle \nabla(y_t|\tilde{\theta}_{t|t}) - \nabla(y_t|\theta_{t|t}), \tilde{\theta}_{t|t} - \theta_{t|t}\rangle,\end{aligned} \tag{B.17}$$

where the second term is non-positive by concavity of the likelihood. We now use the fact that $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ for any $a, b \in \mathbb{R}^d$, which follows from $\|a - b\|^2 = \langle a - b, a - b\rangle = \|a\|^2 + \|b\|^2 - 2\langle a, b\rangle \geq 0$ and reordering. Filling in $a = P_t^{1/2}(\tilde{\theta}_{t|t-1} - \theta_{t|t-1})$ and $b = P_t^{1/2}(\tilde{\theta}_{t|t} - \theta_{t|t})$ and also using concavity of the second term yields

$$\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} \leq \frac{1}{2}\|\tilde{\theta}_{t|t-1} - \theta_{t|t-1}\|^2_{P_t} + \frac{1}{2}\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} - \alpha_t\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2, \tag{B.18}$$

from which it straightforwardly follows that

$$\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t + 2\alpha_t I_K} \leq \|\tilde{\theta}_{t|t-1} - \theta_{t|t-1}\|^2_{P_t}. \tag{B.19}$$

This proves that under Assumptions 1-5, including concavity of the likelihood, we have that the ProPar update is non-expansive with respect the $\|\cdot\|_{P_t}$ norm. The inequality is strict in the case of a strongly concave density ($\alpha_t > 0$) and if $\theta_{t|t} \neq \tilde{\theta}_{t|t}$.

For the second result, we take the derivative of the first-order condition with respect to $\theta_{t|t-1}$. Assuming that the log likelihood is twice differentiable, we obtain

$$H_t\nabla^2(y_t|\theta_{t|t})\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}} = \frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}} - I_K, \tag{B.20}$$

which may be rearranged to yield

$$\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}} = [P_t - \nabla^2(y_t|\theta_{t|t})]^{-1}P_t, \tag{B.21}$$

whereby the existence of $[P_t - \nabla^2(y_t|\theta_{t|t})]^{-1}$ is guaranteed by the twice differentiability assumption and the strict concavity of the regularized likelihood in Assumption 2. This is because under these assumptions the second-order condition reads

$$\nabla^2(y_t|\theta_{t|t}) - P_t \prec O_{K\times K}, \tag{B.22}$$

where $O_{K\times K}$ the $K \times K$ zero matrix and $\prec$ indicates that the right-hand side minus the left-hand side yields a positive definite matrix. Therefore, $P_t - \nabla^2(y_t|\theta_{t|t}) \succ O_{K\times K}$ is positive definite and invertible. As a result, we have that $\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}}$ is the product of two (symmetric) positive definite matrices, such that its smallest eigenvalue is strictly larger than 0. To derive an upper bound for the eigenvalues of $\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}}$ we rewrite (B.21) as follows

$$\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}} = I_K - [P_t - \nabla^2(y_t|\theta_{t|t})]^{-1}(-\nabla^2(y_t|\theta_{t|t})), \tag{B.23}$$

where $(-\nabla^2(y_t|\theta_{t|t}))$ is positive semi-definite by Assumption 5, such that $[P_t - \nabla^2(y_t|\theta_{t|t})]^{-1}(-\nabla^2(y_t|\theta_{t|t}))$ has non-negative eigenvalues. It follows that $\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}}$ has maximum eigenvalue 1. Note that in the case of strict concavity of the log likelihood, we have that the second term has eigenvalues strictly larger than 0, such that the maximum eigenvalue of $\frac{\partial\theta_{t|t}}{\partial\theta'_{t|t-1}}$ is strictly less than 1.

B.5 Lemma 2: Prediction-to-prediction stability

The update-to-prediction mapping from time t to $t+1$ can be written as

$$\begin{aligned} \|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{P_t} = \|\Phi(\theta_{t|t} - \tilde{\theta}_{t|t})\|^2_{P_t} &= -\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t - \Phi' P_t \Phi} + \|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t} && \text{(B.24)} \\ &\leq -\lambda_{\min}(P_t - \Phi' P_t \Phi)\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2 + \|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t} && \text{(B.25)} \\ &\leq \varepsilon_{1,t}\|\theta_{t|t} - \tilde{\theta}_{t|t}\|^2_{P_t}, && \text{(B.26)} \end{aligned}$$

where the second line uses that $\lambda_{\min}(P_t - \Phi' P_t \Phi) \geq 0$ by positive semi-definiteness of $P_t - \Phi' P_t \Phi$, while the last line uses $-\|\cdot\|^2 \leq -\lambda_{\max}(P_t)^{-1}\|\cdot\|^2_{P_t}$. Here $\varepsilon_{1,t}$ is given by

$$\varepsilon_{1,t} = \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t)}. \tag{B.27}$$

By positive definiteness of P_t it follows that $\Phi' P_t \Phi$ is positive semi-definite due to its quadratic form. Therefore, we have that $0 \leq \lambda_{\max}(\Phi' P_t \Phi) = \lambda_{\max}(P_t - (P_t - \Phi' P_t \Phi)) \leq \lambda_{\max}(P_t) + \lambda_{\max}(-(P_t - \Phi' P_t \Phi)) = \lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi) \leq \lambda_{\max}(P_t)$, such that $\varepsilon_{1,t} \in [0,1]$. If $P_t - \Phi' P_t \Phi$ is positive definite, we have that $\varepsilon_{1,t} \in [0,1)$.

In addition, we can write the result of Lemma 1 as

$$(1 + \frac{2\alpha_t}{\lambda_{\max}(P_t)})\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} \leq \|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t + 2\alpha_t I_K} \leq \|\tilde{\theta}_{t|t-1} - \theta_{t|t-1}\|^2_{P_t}, \tag{B.28}$$

which yields

$$\|\tilde{\theta}_{t|t} - \theta_{t|t}\|^2_{P_t} \leq \varepsilon_{2,t}\|\tilde{\theta}_{t|t-1} - \theta_{t|t-1}\|^2_{P_t}, \tag{B.29}$$

where $\varepsilon_{2,t}$ is given as

$$\varepsilon_{2,t} = \frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + 2\alpha_t}. \tag{B.30}$$

Clearly, we have that $\varepsilon_{2,t} \in (0, 1]$ if $\alpha_t \geq 0$ and $\varepsilon_{2,t} \in (0, 1)$ if $\alpha_t > 0$.

Combining (B.26) and (B.29), we obtain

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{P_t} \leq \kappa_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{P_t}, \tag{B.31}$$

where κ_t is given as

$$\begin{aligned} \kappa_t = \varepsilon_{1,t}\varepsilon_{2,t} &= \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t)} \frac{\lambda_{\max}(P_t)}{\lambda_{\max}(P_t) + 2\alpha_t} && \text{(B.32)} \\ &= \frac{\lambda_{\max}(P_t) - \lambda_{\min}(P_t - \Phi' P_t \Phi)}{\lambda_{\max}(P_t) + 2\alpha_t}, && \text{(B.33)} \end{aligned}$$

where if either $\alpha_t > 0$ or $P_t - \Phi' P_t \Phi$ positive definite we have that $\kappa_t \in [0, 1)$, which concludes the proof.

B.6 Theorem 1: Invertibility

By assumption there exists a $\bar{P}$ such that we have for all P_t that $\kappa_t P_t \prec \rho_t \bar{P} \preceq P_t$ for some $\rho_t > 0$. This condition implies that the prediction-to-prediction mapping from time t to $t+1$ is strictly contracting in the norm $\|\cdot\|_{\rho_t \bar{P}}$. To see this, we may write

$$\begin{aligned} \|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{\rho_t \bar{P}} \leq \|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{P_t} &\leq \kappa_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{P_t} && \text{(B.34)} \\ &= -\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\rho_t \bar{P} - \kappa_t P_t} + \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\rho_t \bar{P}} && \text{(B.35)} \\ &\leq -\lambda_{\min}(\rho_t \bar{P} - \kappa_t P_t)\|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2 + \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\rho_t \bar{P}} && \text{(B.36)} \\ &\leq \delta_t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\rho_t \bar{P}}, && \text{(B.37)} \end{aligned}$$

where δ_t is given as

$$\delta_t = \frac{\lambda_{\max}(\rho_t \bar{P}) - \lambda_{\min}(\rho_t \bar{P} - \kappa_t P_t)}{\lambda_{\max}(\rho_t \bar{P})}. \tag{B.38}$$

Due to the condition $\rho_t \bar{P} - \kappa_t P_t \succeq \rho_t A \succ 0$, we obtain that $\delta_t \in [0, \delta]$, where δ is given as

$$\delta = \frac{\lambda_{\max}(\rho_t \bar{P}) - \lambda_{\min}(\rho_t A)}{\lambda_{\max}(\rho_t \bar{P})} = \frac{\lambda_{\max}(\bar{P}) - \lambda_{\min}(A)}{\lambda_{\max}(\bar{P})}, \tag{B.39}$$

where due to positive definiteness of $\bar{P}$ and A we have that $\delta \in (0, 1)$.

It now follows that

$$\|\theta_{t+1|t} - \tilde{\theta}_{t+1|t}\|^2_{\bar{P}} \leq \delta \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\bar{P}}, \tag{B.40}$$

such that every prediction-to-prediction mapping is now strictly contracting in a common norm $\| \cdot \|^2_{\bar{P}}$ with at least strength of contraction $\delta \in (0, 1)$. Therefore we may pick any $c \in (1, \frac{1}{\delta})$ and obtain that

$$\lim_{t \to \infty} c^t \|\theta_{t|t-1} - \tilde{\theta}_{t|t-1}\|^2_{\bar{P}} \to 0, \tag{B.41}$$

such that differences due to initialization disappear exponentially fast almost surely. By norm equivalence it follows that this difference convergences to 0 in any norm.

B.7 Proposition 3: Local KL improvement of the ProPar update

First note that if $\theta_{t|t} \neq \theta_{t|t-1}$, then we must have that $\log p(y_t|\theta_{t|t}) > \log p(y_t|\theta_{t|t-1})$. This follows directly from the definition of $\theta_{t|t}$ as the maximizer of the regularized log-likelihood contribution $f(\theta|y_t, \theta_{t|t-1}, P_t) \;:=\; \log p(y_t|\theta) - \frac{1}{2}\left\|\theta - \theta_{t|t-1}\right\|^2_{P_t}$. For the discrete case, i.e. $\Pr(y = y_t|\theta^0_t) > 0$, we may therefore pick $\delta = 0$ such that $\mathcal{Y} = y_t$ and obtain the desired result.

In the continuous case, we may use the assumed continuity of the postulated density in terms of y to extend the improvement at y_t to a neighbourhood of y_t. Namely, by continuity, $\forall \varepsilon > 0$ with $0 < \varepsilon < \log p(y_t|\theta_{t|t}) - \log p(y_t|\theta_{t|t-1})$, we have that $\exists \delta > 0$ such that $\forall y \in \mathcal{Y} := \{y \in \text{Dom}(y) | \; \|y - y_t\|^2 \leq \delta \; \}$, we have that $\log p(y|\theta_{t|t}) - \log p(y|\theta_{t|t-1}) \geq \log p(y_t|\theta_{t|t}) - \log p(y_t|\theta_{t|t-1}) - \varepsilon > 0$. In addition, we have that $\forall \delta > 0$, that $\Pr(y \in \mathcal{Y}|\theta^0_t) := \int_{\mathcal{Y}} p_0(y|\theta^0_t) \mathrm{d}y > 0$ (otherwise y_t could not have occurred), which completes the proof.

B.8 Lemma 3: Contractive and expansive forces

We write the first-order condition of the ProPar update as follows

$$H_t^{-1/2}(\theta_{t|t} - \theta_{t|t-1}) = H_t^{1/2} \nabla(y_t|\theta_{t|t}), \tag{B.42}$$

adding $H_t^{1/2}\nabla(y_t|\theta^\star_t) - H_t^{-1/2}\theta^\star_t$ to both sides and rearranging gives

$$H_t^{-1/2}(\theta_{t|t} - \theta^\star_t) + H_t^{1/2}(\nabla(y_t|\theta^\star_t) - \nabla(y_t|\theta_{t|t})) = H_t^{-1/2}(\theta_{t|t-1} - \theta^\star_t) + H_t^{1/2}\nabla(y_t|\theta^\star_t). \tag{B.43}$$

Taking the quadratic norm yields

$$\|\theta_{t|t} - \theta_t^\star\|_{P_t}^2 + \|\nabla(y_t|\theta_t^\star) - \nabla(y_t|\theta_{t|t})\|_{H_t}^2 + 2\langle\nabla(y_t|\theta_t^\star) - \nabla(y_t|\theta_{t|t}), \theta_{t|t} - \theta_t^\star\rangle \tag{B.44}$$

$$= \|\theta_{t|t-1} - \theta_t^\star\|_{P_t}^2 + \|\nabla(y_t|\theta_t^\star)\|_{H_t}^2 + 2\langle\nabla(y_t|\theta_t^\star), \theta_{t|t-1} - \theta_t^\star\rangle. \tag{B.45}$$

We now take the expectation over y_t with respect to the DGP on both sides and use that $\mathop{\mathbb{E}}_{y_t}[\nabla(y_t|\theta_t^\star)] = 0$ by Assumption 6 and that $\|\nabla(y_t|\theta_t^\star) - \nabla(y_t|\theta_{t|t})\|_{H_t}^2 \geq 0$ to obtain

$$\mathop{\mathbb{E}}_{y_t}[\|\theta_{t|t} - \theta_t^\star\|_{P_t}^2] \leq \|\theta_{t|t-1} - \theta_t^\star\|_{P_t}^2 + 2\mathop{\mathbb{E}}_{y_t}[\langle\nabla(y_t|\theta_{t|t}) - \nabla(y_t|\theta_t^\star), \theta_{t|t} - \theta_t^\star\rangle] + \mathop{\mathbb{E}}_{y_t}[\|\nabla(y_t|\theta_t^\star)\|_{H_t}^2], \tag{B.46}$$

which concludes the proof.

B.9 Theorem 2: Contraction to the NDR

The statements follow directly from substituting the expressions in Assumptions 8a and 8b in Lemma 3. Note that Assumption 5 (concave log-likelihood function) is not needed for deriving the particular expression in Lemma 3; it is only used to determine the signs of the components on the right-hand side.