The Efficiency Gap

Dimitriadis, Timo; Fissler, Tobias; Ziegel, Johanna F.

DOI:
10.14146

Published: 01/01/2020

Document Version
Unknown

The Efficiency Gap

Timo Dimitriadis* Tobias Fissler† Johanna F. Ziegel‡

October 28, 2020

Abstract. Parameter estimation via M- and Z-estimation is broadly considered to be equally powerful in semiparametric models for one-dimensional functionals. This is due to the fact that, under sufficient regularity conditions, there is a one-to-one relation between the corresponding objective functions – strictly consistent loss functions and oriented strict identification functions – via integration and differentiation. When dealing with multivariate functionals such as multiple moments, quantiles, or the pair (Value at Risk, Expected Shortfall), this one-to-one relation fails due to integrability conditions: Not every identification function possesses an antiderivative. The most important implication of this failure is an *efficiency gap*: The most efficient Z-estimator often outperforms the most efficient M-estimator, implying that the semiparametric efficiency bound cannot be attained by the M-estimator in these cases. We show that this phenomenon arises for pairs of quantiles at different levels and for the pair (Value at Risk, Expected Shortfall), where we illustrate the gap through extensive simulations.

Keywords: Efficient estimation; Expected Shortfall; M-estimation; Quantiles; Semiparametric efficiency bound, Semiparametric regression; Value at Risk; Z-estimation
MSC2020 classes: 62F10; 62F12; 62J02; 62M10

1. Introduction

Given some real-valued response variable Y and some p-dimensional vector of covariates X, one is often interested in modelling the effect of the covariates on the response variable through regression models. E.g., one might be interested in the (average) effect of education on the lifetime income of individuals. The classical (mean) regression technique captures the *average* effect by modelling the *expectation* of the conditional distribution of Y given X, $F_{Y|X}$. However, researchers are often interested in different properties of this conditional distribution, e.g., in low quantiles if attention is focused on the effect of education specifically for low income respondents (Angrist et al., 2006). This can be facilitated through quantile regression (Koenker and Bassett, 1978), where one parametrically models the quantile of the conditional distribution $F_{Y|X}$.

*Heidelberg University, Alfred Weber Institute of Economics, 69115 Heidelberg, Germany and Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany, e-mail: timo.dimitriadis@awi.uni-heidelberg.de

†WU Vienna University of Economics and Business, Department of Finance, Accounting and Statistics, Welthandelsplatz 1, 1020 Vienna, Austria, e-mail: tobias.fissler@wu.ac.at

‡University of Bern, Department of Mathematics and Statistics, Institute of Mathematical Statistics and Actuarial Science, Alpeneggstrasse 22, 3012 Bern, Switzerland, e-mail: johanna.ziegel@stat.unibe.ch

More generally, one is interested in a certain statistical *functional* Γ of the conditional distribution $F_{Y|X}$, where the functional maps a (conditional) distribution to a real-valued outcome. The functional of interest varies among disciplines: E.g., quantitative risk management is particularly interested in models for risk measures such as conditional variances, quantiles (Value at Risk, VaR), expectiles and Expected Shortfall (ES) (Bollerslev, 1986; Engle and Manganelli, 2004; Efron, 1991; Patton et al., 2019). Epidemiological forecasts, of particular interest due to the recent spread of COVID-19, often focus on prediction intervals, which commonly consist of two (conditional) quantiles (Bracher et al., 2020).

It is common practice to model the functional $\Gamma(F_{Y|X})$ as some parametric model $m(X, \theta)$ of X, with a parameter $\theta \in \Theta \subseteq \mathbb{R}^q$, i.e.

$$\Gamma(F_{Y|X}) = m(X, \theta_0), \qquad \text{for some unique } \theta_0 \in \Theta. \tag{1.1}$$

The specification (1.1) is commonly referred to as *semiparametric*: Even though the model itself is parametric, it does not specify the full conditional distribution $F_{Y|X}$, but only a functional thereof (Newey, 1990; Bickel et al., 1998).

While standard approaches often model every functional of interest separately, joint semiparametric models for *multivariate* (or vector-valued) functionals have desirable advantages in many instances: A joint treatment of two quantile levels is desirable e.g. for prediction intervals, and it can impede quantile crossings (Gourieroux and Jasiak, 2008; White et al., 2015). Even more fundamental, there are cases where univariate modeling is infeasible such as for the variance, ES and Range Value at Risk (RVaR), where suitable loss or identification functions do not exist (Osband, 1985; Weber, 2006; Wang and Wei, 2020). However, such objective functions exist for an appropriate multivariate functional; see Fissler and Ziegel (2016) for the pair (VaR, ES), Osband (1985) for the pair (mean, variance), and Fissler and Ziegel (2019b) for the triplet of the RVaR with two quantiles. These examples motivate our consideration of (efficient) joint estimation of such multivariate models.

A key task in statistics and econometrics is to come up with a 'good' estimator $\widehat{\theta}_T$ of the true (but unknown) parameter θ_0, given data (Y_t, X_t) for $t = 1, \dots, T$. Basic desirable criteria of such an estimator are consistency, meaning that θ_0 is the probability limit of $\widehat{\theta}_T$, and asymptotic normality, i.e., the stabilising transformation $\sqrt{T}(\widehat{\theta}_T - \theta_0)$ approaches a Gaussian distribution. Given that asymptotic normality holds, one commonly favours an *efficient* estimator, i.e. an estimator with an associated covariance matrix which is as small as possible. Besides more accurate estimates, this allows for more powerful inference through tests and confidence intervals.

For the most standard situation of linear mean regression (i.e., $\Gamma(F_{Y|X}) = \mathbb{E}[Y|X]$, and $m(X, \theta) = X^\intercal \theta$), one often employs the ordinary-least-squares (OLS) estimator, $\widehat{\theta}_{\text{OLS},T} = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T \big(Y_t - X_t^\intercal \theta\big)^2$, or a related closed-form solution thereof. The OLS estimator is a special instance of an M-estimator (Huber, 1967; Newey and McFadden, 1994),

$$\widehat{\theta}_{M,T} = \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T \rho_t\big(Y_t, m(X_t, \theta)\big), \tag{1.2}$$

based on some (possibly time-varying) loss functions ρ_t, which are the key ingredient of an M-estimator. A core condition on ρ_t for the consistency of $\widehat{\theta}_{M,T}$ is that

$$\mathbb{E}\big[\rho_t\big(Y_t, m(X_t, \theta_0)\big)\big] < \mathbb{E}\big[\rho_t\big(Y_t, m(X_t, \theta)\big)\big] \qquad \forall \theta \neq \theta_0 \quad \forall t \in \mathbb{N}, \tag{1.3}$$

which we call *strict model-consistency* of ρ_t for m.

A standard alternative to M-estimation is (zero) Z-estimation or, more generally, generalized method of moments (GMM) estimation (Hansen, 1982; Newey and McFadden, 1994), given by

$$\widehat{\theta}_{\mathrm{GMM},T} = \arg\min_{\theta\in\Theta} \left(\frac{1}{T} \sum_{t=1}^{T} \psi_t(Y_t, X_t, \theta) \right)^{\intercal} P_T \left(\frac{1}{T} \sum_{t=1}^{T} \psi_t(Y_t, X_t, \theta) \right), \tag{1.4}$$

based on a symmetric and positive-definite weighting matrix P_T and on some (possibly time-varying) s-dimensional functions $\psi_t(Y_t, X_t, \theta)$. The latter are often called *moment conditions* or *identification functions* for θ_0 satisfying the *strict unconditional identification* condition

$$\Big(\mathbb{E}\big[\psi_t(Y_t, X_t, \theta)\big] = 0 \quad \iff \quad \theta = \theta_0\Big) \qquad \forall \theta \in \Theta \quad \forall t \in \mathbb{N}. \tag{1.5}$$

For the consideration of *efficient* GMM-estimation, it suffices to restrict attention to the *exactly identified* case of $s = q$, implying as many moment conditions as model parameters, and on $P_T = I_q$, the identity matrix on $\mathbb{R}^q$; see Theorem 3.1 and Remark 3.3 for details. Motivated by condition (1.5), the resulting estimator is called Z-estimator, and is given by

$$\widehat{\theta}_{Z,T} = \arg\min_{\theta\in\Theta} \left\| \frac{1}{T} \sum_{t=1}^{T} \psi_t(Y_t, X_t, \theta) \right\|^2. \tag{1.6}$$

Similar to (1.3), the strict unconditional identification (1.5) is a core condition for the consistency of the Z-estimator $\widehat{\theta}_{Z,T}$ and the choice of ψ_t is the key ingredient of Z-estimation, which also governs its efficiency.

Loosely speaking, the identification condition (1.5) can often be interpreted as the first-order condition(s) of the minimisation condition (1.3), if ψ_t equals the derivative of ρ_t with respect to θ. This illustrates the coherence of M- and Z-estimation through the often closely-related choices of ρ_t and ψ_t, also in terms of estimation efficiency, where the general task is to find (possibly time-varying) choices of ρ_t and ψ_t which result in most efficient estimators. See e.g. Gourieroux et al. (1984) and Komunjer and Vuong (2010b,a) for examples of efficient estimation of semiparametric models for the mean and quantiles. Upon 'matching ρ_t and ψ_t,' one ends up with the general observation that the efficiency bounds for M- and Z-estimation coincide for the estimation of *univariate* functionals.

As a main contribution of this article, we illustrate that this equivalence in terms of efficient estimation does not hold for many examples of semiparametric models for vector-valued functionals. Instead, the classes of (consistent) M-estimators are considerably smaller than the corresponding classes of (consistent) Z-estimators, generating an *efficiency gap* between M- and Z-estimation: The most efficient M-estimator does not attain the Z-estimation efficiency bound. We establish the Z-estimation efficiency bound for joint models for mean and second moment, mean and variance models, VaR and ES, and for the double quantile model. While the efficiency gap does not substantiate in the first two cases (Subsection 4.1), we derive conditions when this gap is present for the latter two cases (Subsections 4.2 and 4.3), both of which recently gained attention in risk management and through prediction intervals. We numerically illustrate these results through a simulation study considering different examples of modelling multivariate functionals through semiparametric (linear) models (Section 5). We anticipate that this gap generalises to joint models for various other vector-valued functionals like multiple expectiles or the interquantile expectation (the trimmed mean, or RVaR), jointly with its quantiles.

The following subsection gives a more technical and detailed overview of the other major contributions of the paper.

1.1. Detailed Overview of the Article

In order to technically formalise the efficiency gap, in Subsection 2.2 we characterise the class of (consistent) M-estimators based on condition (1.3) through relating them to the classes of strictly consistent loss functions from the theory of forecast evaluation. By virtue of (1.1), Theorem 2.5 essentially establishes that a loss function ρ is strictly model-consistent for m if and only if it is strictly *consistent* for the functional Γ, meaning that $\mathbb{E}_{Z\sim F}[\rho(Z,\Gamma(F))] < \mathbb{E}_{Z\sim F}[\rho(Z,\xi)]$ for all $\xi \neq \Gamma(F)$ and for all F in a sufficiently large class consisting of the conditional distributions $F_{Y_t|X_t}$. While this latter condition is closely related to (1.3), it is crucially different in the sense that the expectation is taken with respect to the *conditional* distribution of Y_t given X_t, whereas in (1.3) it is taken with respect to *joint* distribution of (X_t, Y_t).

Since there are well-understood characterisation results for strictly consistent losses (Gneiting, 2011a; Fissler and Ziegel, 2016), Theorem 2.5 lifts these results to a novel characterisation of the classes of consistent M-estimators of the form (1.2), thereby generalising the results of Gourieroux et al. (1984) for conditional mean models and Komunjer and Vuong (2010b) for conditional quantile models by considering general (vector-valued) elicitable functionals, i.e., functionals possessing a strictly consistent loss function.

While an equivalent characterisation result for (consistent) Z-estimators is generally desirable, it is not required for theoretically establishing the efficiency gap, nor is it as easily available; see e.g. Komunjer (2012) among others for details. While we make some progress in this direction in Appendix B, establishing an analogon of Theorem 2.5 and an exhaustive characterisation of the class of consistent Z-estimators of the form (1.6) seems to be out of reach, even though a novel characterisation for strict identification functions φ for a general functional Γ can be provided (Theorem B.1).

Following e.g. Newey (1993) among many others, we focus on moment functions of the form

$$\psi_t(Y_t, X_t, \theta) = A_t(X_t, \theta)\varphi\big(Y_t, m(X_t, \theta)\big), \tag{1.7}$$

where $A_t(X_t, \theta)$ is a suitable (possibly time-varying) *instrument matrix*, usually satisfying some full rank condition, and φ is a *strict identification function* for Γ, sometimes also referred to as *conditional moment restrictions*. Focusing on Z-estimators of the form (1.7) facilitates considerations of efficient estimation due to the well-known form of the *efficient instruments* of Hansen (1985), Chamberlain (1987), Newey (1993), and (subject to regularity conditions), their relation to the *semiparametric efficiency bound* in the sense of Stein (1956). Our Theorem 3.1 shows that this efficient instrument choice is not only sufficient, but also necessary, for efficient Z-estimation. Furthermore, relating the efficient instrument choice to the classes of (consistent) M-estimators is straight-forward through differentiation of the feasible (strictly consistent) loss functions due to the following consideration.

If the functional of interest Γ is univariate, the two approaches of M- and Z-estimation are equally powerful due to the intimate relationship of strictly consistent losses ρ and strict identification functions φ for Γ: Under suitable regularity and smoothness conditions, an oriented strict identification function φ gives rise to a strictly consistent loss ρ via an integral construction, known as "Osband's principle"; see Osband (1985), Gneiting (2011a), Steinwart et al. (2014), and Fissler and Ziegel (2016). Vice versa, differentiation of ρ yields an identification function φ. The most prominent example are the functions $\rho(y,\xi) = \frac{1}{2}(y-\xi)^2$ and $\varphi(y,\xi) = y - \xi$, for the mean functional.

However, this convenient one-to-one relationship breaks down when Γ is multivariate: Due to integrability conditions (Königsberger, 2004, p. 185), not every identification function possesses an antiderivative. This recent observation implies a gap between consistent loss functions and identification functions, and gives rise to a *gap* between the class of (consistent) M- and Z-estimators if the underlying functional Γ is multivariate, even though a strict counterpart of Theorem 2.5 is illusive.

The paper is organised as follows. Section 2 introduces the notation, and characterises and relates the classes of strictly consistent loss and identification functions. Some additional results are deferred to Appendix B. Section 3 considers efficient M- and Z-estimation of general semi-parametric models and attainability of the semiparametric efficiency bound. In Section 4, we consider several examples of semiparametric models for vector-valued functionals, demonstrating the efficiency gap, which is numerically illustrated in Section 5. Section 6 concludes with a discussion and outlook. Appendix A describes further implications of the gap for equivariant estimation. Appendix C provides some technical background about the efficient Z-estimator for the double quantile model. We report additional simulation results in Appendix E and defer all proofs to Appendix F.

2. Loss and identification functions

2.1. Notation and Setting

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a non-atomic, complete probability space where all random objects are defined. We equip all Euclidean spaces with the induced Borel σ-algebras. We consider a time series $(Z_t)_{t\in\mathbb{N}}$ where $Z_t = (Y_t, X_t)$. Adopting classical notation in regression, Y_t are real-valued response variables and X_t are $\mathbb{R}^p$-valued regressors. We introduce the classes of possible random variables $\mathcal{Y} \subseteq \mathcal{L}^0(\Omega;\mathbb{R})$, $\mathcal{X} \subseteq \mathcal{L}^0(\Omega;\mathbb{R}^p)$, $\mathcal{Z} \subseteq \mathcal{L}^0(\Omega;\mathbb{R}\times\mathbb{R}^p)$, and assume that $Y_t \in \mathcal{Y}$, $X_t \in \mathcal{X}$ and $Z_t \in \mathcal{Z}$ for all $t \in \mathbb{N}$. Let $\mathcal{F}_\mathcal{Z} = \{F_Z \mid Z \in \mathcal{Z}\}$ and $\mathcal{F}_\mathcal{X} = \{F_X \mid X \in \mathcal{X}\}$ be the classes of joint distributions and marginal distributions, respectively. Additionally, let $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ be a collection of one-dimensional distributions such that for any $(Y,X) \in \mathcal{Z}$ a regular version of the conditional distribution $F_{Y|X}$ is an element of $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ almost surely.[1] Also assume that $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ contains only regular versions of conditional distributions $F_{Y|X}$ for $(Y,X) \in \mathcal{Z}$. We will identify cumulative distribution functions with their corresponding measures where convenient. In this section, we conveniently drop the time index t when the time dependence is inessential.

Let $\Gamma\colon \mathcal{F}_{\mathcal{Y}|\mathcal{X}} \to \Xi \subseteq \mathbb{R}^k$ be some k-dimensional functional of the conditional distribution of Y given X.[2] Throughout the entire article we shall tacitly assume measurability of all functions introduced. Let $\Theta \subseteq \mathbb{R}^q$ be some parameter space with non-empty interior, $\mathrm{int}(\Theta)$, and $m\colon \mathbb{R}^p \times \Theta \to \Xi$ a parametric model for the functional Γ. We say that m is differentiable if for all $X \in \mathcal{X}$, all components $m_j(X,\cdot)$, $j \in \{1,\dots,k\}$ are differentiable almost surely and the gradient of such a component with respect to θ is denoted by the column vector $\nabla_\theta m_j(X,\theta) \in \mathbb{R}^q$. We shall work under the following assumption of a correctly specified model with a unique parameter.

Assumption (1). For all $Z = (Y,X) \in \mathcal{Z}$ there is a unique parameter $\theta_0 = \theta_0(F_Z) \in \mathrm{int}(\Theta)$ such that almost surely

$$m(X,\theta_0) = \Gamma(F_{Y|X}) \tag{2.1}$$

[1] Recall that $F_{Y|X}$ is unique almost surely, or equivalently, $F_{Y|X=x}$ is unique for $\mathbb{P}\circ X^{-1}$-almost all $x \in \mathbb{R}^p$ (that is for all x in some set $B \subseteq \mathbb{R}^p$ such that $\mathbb{P}(X \in B) = 1$). Also note that $\mathcal{F}_{\mathcal{X}|\mathcal{Y}}$ is not unique, but generally has multiple modifications or versions in the sense that for any $(Y,X) \in \mathcal{Z}$ the null set $A_0 \in \mathcal{A}$ of $\omega \in \Omega$ such that $F_{Y|X}(\cdot,\omega) \notin \mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ might be different.

[2] In particular, we assume that Γ is single-valued and not set-valued. The set-valued case requires a technical treatment of its own; see e.g. Fissler et al. (2020). Many functionals of interest, such as moments or expectiles, are single-valued valued *per se*. For α-quantiles, we assume that the conditional distributions $F_{Y|X=x}$ are strictly increasing – at least locally at their α-quantile – such that the quantiles are singletons.

Remark 2.1. Assumption (1) induces a functional $\theta_0 \colon \mathcal{F}_\mathcal{Z} \to \Theta$ by mapping a distribution $F_Z \in \mathcal{F}_\mathcal{Z}$ to the unique $\theta_0(F_Z) \in \operatorname{int}(\Theta)$ such that (2.1) is satisfied. However, to save notation and to be coherent with the traditional econometric notation, the dependence on the joint distribution is regularly dropped and we merely write θ_0.

Remark 2.2. Interestingly, under the additional surjectivity assumption that for any $\theta \in \operatorname{int}(\Theta)$ there is some $Z = (Y, X) \in \mathcal{Z}$ such that $\theta_0(F_Z) = \theta$, the uniqueness stipulated in Assumption (1) implies that that for any $\theta, \theta' \in \operatorname{int}(\Theta)$: If $m(x, \theta) = m(x, \theta')$ for all $x \in \mathbb{R}^p$ then $\theta = \theta'$. This injectivity property implies that the model m uniquely identifies the parameter.

We will dispense with a strong stationarity assumption on the time series $(Z_t)_{t\in\mathbb{N}}$. What we still need, however, is a *semiparametric stationarity* assumption:

Assumption (2). Let Assumption (1) hold. Assume there is a unique $\theta_0 \in \operatorname{int}(\Theta)$ such that $\theta_0 = \theta_0(F_{Z_t})$ for all $t \in \mathbb{N}$.

Together with Assumption (1), the time-constancy of the parameter $\theta_0(F_{Z_t})$ in Assumption (2) implies that the functional under consideration $\Gamma(F_{Y_t|X_t})$ is time-independent, while the remaining conditional distributions may vary. We do not explicitly impose any dependence conditions on the process, however, we assume consistency and asymptotic normality of the considered estimators in Section 3 and 4, which implicitly rely on a restricted dependence in the form of ergodicity or mixing-conditions, see e.g. White (2001) for details.

2.2. Elicitability and (un-)conditional consistency

The goal of this section is to obtain a clear understanding of the class of consistent M-estimators of the form in (1.2). This amounts to determining possible loss functions ρ which are the crucial building block in the M-estimator (1.2). More precisely, we determine the class of M-estimators – or equivalently loss functions – which satisfy the necessary condition of strict *model-consistency* defined in (1.3) by relating this to well-studied characterisation results for strictly consistent losses for Γ from the forecast evaluation literature.

The notion of a *strictly consistent* (Murphy and Daan, 1985) loss function is a crucial concept in the literature on forecast evaluation, since it incentivises truthful reports. Making use of a similar decision-theoretic terminology as in Gneiting (2011a) or Fissler and Ziegel (2016), let $\mathcal{F}$ be some generic class of probability distributions on $\mathbb{R}$, which is our *observation domain*, and $\Xi \subset \mathbb{R}^k$ our *action domain*. (Occasionally, we shall also consider the situation where the parameter space Θ is the action domain. Then the respective definitions apply *mutatis mutandis*.) Moreover, let $\Gamma \colon \mathcal{F} \to \Xi$ be the functional of interest. We call a function $a \colon \mathbb{R} \times \Xi \to \mathbb{R}$ $\mathcal{F}$-integrable if $a(\cdot, \xi)$ is F-integrable for all $F \in \mathcal{F}$ and for all $\xi \in \Xi$. Then we write $\bar{a}(F, \xi) := \int a(y, \xi)\, \mathrm{d}F(y)$ for $\xi \in \Xi$, $F \in \mathcal{F}$.

Definition 2.3 (Consistency and elicitability)**.** A loss function $\rho \colon \mathbb{R} \times \Xi \to \mathbb{R}$ is called $\mathcal{F}$-*consistent* for a functional $\Gamma \colon \mathcal{F} \to \Xi$ if it is $\mathcal{F}$-integrable and if

$$\bar{\rho}\big(F, \Gamma(F)\big) \le \bar{\rho}(F, \xi) \qquad \text{for all } F \in \mathcal{F},\ \xi \in \Xi\,. \tag{2.2}$$

If equality in (2.2) implies that $\xi = \Gamma(F)$ then the loss function is called *strictly $\mathcal{F}$-consistent* for Γ. A functional $\Gamma \colon \mathcal{F} \to \Xi$ is *elicitable* if there is a strictly $\mathcal{F}$-consistent loss function for it.

For example, the squared loss $\rho(y, \xi) = (y - \xi)^2$ is strictly $\mathcal{F}$-consistent for the mean, if $\mathcal{F}$ contains only distributions with a finite second moment. More generally, under certain richness conditions on $\mathcal{F}$, one can show that ρ is strictly $\mathcal{F}$-consistent for the mean if and only if it is a *Bregman loss* $\rho(y,\xi) = -\phi(\xi) + \phi'(\xi)(\xi - y) + \kappa(y)$ where ϕ is a strictly convex function on $\mathbb{R}$ with subgradient ϕ' and κ is $\mathcal{F}$-integrable (Savage, 1971; Gneiting, 2011a). Likewise, if $\mathcal{F}$ contains only distributions with a unique α-quantile, a loss is strictly $\mathcal{F}$-consistent for the α-quantile with $\alpha \in (0,1)$, if and only if ρ is a generalised piecewise linear loss functions $\rho(y,\xi) = \mathbb{1}\{y \le \xi\}(g(y) - g(\xi)) + \alpha g(\xi) + \kappa(y)$, where g is strictly increasing, and $\mathbb{1}\{\cdot \le \xi\}g(\cdot)$ and κ are $\mathcal{F}$-integrable (Gneiting, 2011b). This class nests the well known pinball (or asymmetric absolute) loss $\rho(y,\xi) = (\mathbb{1}\{y \le \xi\} - \alpha)(y - \xi)$.

As outlined in the introduction, in semiparametric M-estimation, we are chiefly interested in the following notion of (unconditional) *model-consistency* which is also accompanied by a conditional version.

Definition 2.4 (Model-consistency)**.** Suppose Assumption (1) holds for the parametric model $m\colon \mathbb{R}^p \times \Theta \to \Xi$ and the functional $\Gamma\colon \mathcal{F}_{\mathcal{Y}|\mathcal{X}} \to \Xi$. Let $\rho\colon \mathbb{R} \times \Xi \to \mathbb{R}$ be a loss function such that $\mathbb{E}\big|\rho\big(Y, m(X,\theta)\big)\big| < \infty$ for all $(Y,X) \in \mathcal{Z}$ and $\theta \in \Theta$.

(i) The loss ρ is *unconditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent* for the model m if

$$\mathbb{E}\big[\rho\big(Y, m(X,\theta_0)\big)\big] \le \mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big] \qquad \text{for all } (Y,X) \in \mathcal{Z},\ \theta \in \Theta\,. \tag{2.3}$$

Moreover, ρ is *strictly unconditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent* for the model m if equality in (2.3) implies that $\theta = \theta_0$.

(ii) The loss ρ is *conditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent* for the model m if

$$\mathbb{E}\big[\rho\big(Y, m(X,\theta_0)\big)\big|X\big] \le \mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big|X\big] \quad \text{a.s.} \quad \text{for all } (Y,X) \in \mathcal{Z},\ \theta \in \Theta\,. \tag{2.4}$$

Moreover, ρ is *strictly conditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent* for the model m if almost sure equality in (2.4) implies that $\theta = \theta_0$.

Whenever we merely speak of (strict) model-consistency, we mean (strict) *unconditional* model consistency. Note that (1.2) approximates the *unconditional* expectation rendering the concept of unconditional model-consistency central for practical purposes of M-estimation. The corresponding conditional notion of consistency can still be practically useful when we have repeated observations of X (e.g. if X consists of categorical variables only), or when we resort to kernel estimation methods. Theoretically, the conditional notion is appealing since it bridges the gap between consistency for Γ according to Definition 2.3 and unconditional model-consistency for m. The following Theorem 2.5 makes this statement formal. We first introduce the following assumptions.

Assumption (3). For all $X \in \mathcal{X}$, the map $m(X,\cdot)\colon \Theta \to \Xi$ is surjective almost surely. For all $(Y,X) \in \mathcal{Z}$ the conditional expectation $\mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big|X\big]$ is continuous in θ almost surely.

Assumption (4). For any $Z \in \mathcal{Z}$ and any $\sigma(X)$-measurable event $A \in \mathcal{A}$ with positive probability $\mathbb{P}(A) > 0$, there is some $\widetilde{Z} \in \mathcal{Z}$ such that $\mathbb{P}\big(\widetilde{Z} \in B\big) = \mathbb{P}\big(Z \in B\,|A\big)$ for all Borel sets $B \subseteq \mathbb{R}^{p+1}$.

While Assumption (3) is obviously a smoothness condition, Assumption (4) is a richness condition on the class of possible data generating processes (DGPs) in $\mathcal{Z}$. Crucially, for $Z = (Y,X)$, A and $\widetilde{Z} = (\widetilde{Y}, \widetilde{X})$ as specified in Assumption (4), it yields that $\mathbb{P}(\widetilde{Y} \in C\,|\widetilde{X}) = \mathbb{P}(Y \in C\,|X, A)$ for all Borel sets $C \subseteq \mathbb{R}$. This, together with Assumption (1), implies that the correctly specified parameter (and hence the semiparametric model) is the same under the distributions $F_{\widetilde{Z}}$ and F_Z; in formulae, $\theta_0(F_{\widetilde{Z}}) = \theta_0(F_Z)$. Recall that in estimation, $\mathcal{Z}$ captures the flexibility about the underlying (in practice unknown) DGP, such that a large $\mathcal{Z}$ is desirable in order to

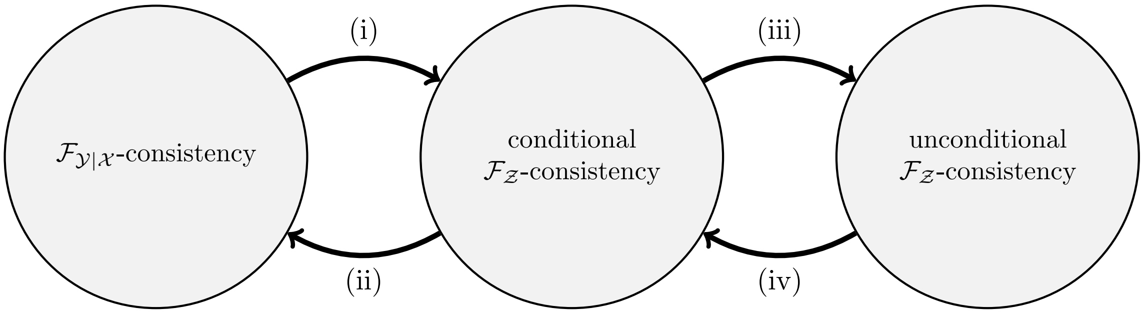

Figure 1.: A visualisation of the four implications established in Theorem 2.5.

obtain an estimation method which is applicable to a wide range of distributions of $Z_t \in \mathcal{Z}$. Thus, Assumption (4) intuitively means that given a certain plausible (and correctly specified) DGP, and given a set $D \subset \mathbb{R}^p$ of possible values for the covariates X which is attained with positive probability, i.e. $\mathbb{P}(X \in D) > 0$, restricting the DGP to these values of covariates must be feasible. E.g., if income, Y, is studied in dependence of years after graduation, X_1, and further covariates $X_2, \ldots, X_p$, one might as well study income of persons at most 5 years after their graduation, $X_1 \leq 5$. Then, in a correctly specified model, the true (but unknown) regression parameter θ_0 remains the same, no matter whether considering the whole population or only persons within 5 years after their graduation.

Theorem 2.5. *Under Assumption (1) the following implications hold for a loss* $\rho\colon \mathbb{R} \times \Xi \to \mathbb{R}$.

(i) *If* ρ *is (strictly)* $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$*-consistent for* Γ *then it is (strictly) conditionally* $\mathcal{F}_{\mathcal{Z}}$*-model-consistent for the model* m.

(ii) *Under Assumption (3) if* ρ *is conditionally* $\mathcal{F}_{\mathcal{Z}}$*-model-consistent for* m*, there is a version* $\widetilde{\mathcal{F}}_{\mathcal{Y}|\mathcal{X}}$ *of* $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ *such that* ρ *is* $\widetilde{\mathcal{F}}_{\mathcal{Y}|\mathcal{X}}$*-consistent for* Γ.

(iii) *If* ρ *is (strictly) conditionally* $\mathcal{F}_{\mathcal{Z}}$*-model consistent for* m *then it is (strictly) unconditionally* $\mathcal{F}_{\mathcal{Z}}$*-model-consistent for* m.

(iv) *Under Assumption (4) if* ρ *is (strictly) unconditionally* $\mathcal{F}_{\mathcal{Z}}$*-model-consistent for* m *then it is also (strictly) conditionally* $\mathcal{F}_{\mathcal{Z}}$*-model-consistent for* m.

Theorem 2.5 provides two main implications; see Figure 1 for a visualisation: First, a combination of (i) and (iii) justifies the usage of strictly $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent losses for Γ in the context of M-estimation. This is well known in the literature, e.g., Gneiting and Raftery (2007, Section 9) describe this under the term “optimum score estimation”. The proofs of (i) and (iii) are straight forward, and for special cases they can be found, e.g., in the proof of Patton et al. (2019, Theorem 1). Second – and more important for our purposes – is the reverse implication, combining (ii) and (iv). It asserts that, under appropriate assumptions, an unconditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent loss is necessarily $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent. Thus, exploiting known characterisation results for $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent losses for many relevant functionals Γ, it constitutes an effective and original bound on the class of consistent M-estimators. Notice that *strictness* of the $\widetilde{\mathcal{F}}_{\mathcal{Y}|\mathcal{X}}$-consistent loss cannot be established in part (ii) of Theorem 2.5. While a stronger version of this result (including the strictness) would be desirable, its lack merely diminishes the applicability of the results since characterisation results for (possibly non-strict) $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent losses are available in the literature and these are almost as strong as the characterisations for strictly consistent losses (Gneiting, 2011a). As of yet, an implication in the direction of the points (ii) and (iv) of Theorem 2.5 has only been provided for special cases; e.g., Komunjer (2005) shows it if Γ is some quantile. Note that for practical or intuitive purposes, the technical distinction between $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ and a version $\widetilde{\mathcal{F}}_{\mathcal{Y}|\mathcal{X}}$ thereof is inessential; see footnote 1 for some remarks.

Remark 2.6. In the light of Assumption (1) and Remark 2.1, one can actually regard a (strictly) unconditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent loss function $\rho\colon \mathbb{R}\times\Xi\to\mathbb{R}$ for a model $m\colon \mathbb{R}^p\times\Theta\to\Xi$ as a (strictly) $\mathcal{F}_{\mathcal{Z}}$-consistent loss function for the functional $\theta_0\colon \mathcal{F}_{\mathcal{Z}}\to\Theta$ of the form $\rho_m\colon \big(\mathbb{R}\times\mathbb{R}^p\big)\times\Theta\to\mathbb{R}$, $\big((y,x),\theta\big)\mapsto\rho_m(y,x,\theta)=\rho\big(y,m(x,\theta)\big)$ in the sense of Definition 2.4. This means ρ_m has the observation domain $\mathbb{R}\times\mathbb{R}^p$ rather than $\mathbb{R}$ and the action domain $\Theta\subseteq\mathbb{R}^q$ rather than $\Xi\subseteq\mathbb{R}^k$. With the same rationale, one can say that $\theta_0\colon \mathcal{F}_{\mathcal{Z}}\to\Theta$ is elicitable if there is a strictly $\mathcal{F}_{\mathcal{Z}}$-consistent loss function of the form $\rho_m\colon (\mathbb{R}\times\mathbb{R}^p)\times\Theta\to\mathbb{R}$.

2.3. (Un-)conditional identifiability

The previous section provides a clear picture of the class of possible consistent M-estimators of the form (1.2). In particular, Theorem 2.5 connects consistent losses for Γ with model-consistent losses for m: The class of consistent M-estimators is a subclass of the collection of consistent loss functions for Γ. The aim of this section is to provide a similar understanding of the class of Z-estimators of the form (1.6). To this end, we strive for a similar characterisation of moment functions ψ satisfying (1.5) and to link them to identification functions for the functional Γ. As a matter of fact, it is only feasible to establish a counterpart of Theorem 2.5 (i) and (ii), see Lemma 2.9. We defer partial results on the other implications to Appendix B.

Definition 2.7 (Identification function and identifiability)**.** An $\mathcal{F}$-integrable map $\varphi\colon \mathbb{R}^d\times\Xi\to\mathbb{R}^\ell$ is an *$\mathcal{F}$-identification function* for a functional $\Gamma\colon \mathcal{F}\to\Xi\subseteq\mathbb{R}^k$ if $\bar\varphi\big(F,\Gamma(F)\big)=0$ for all $F\in\mathcal{F}$. If additionally $\bar\varphi(F,\xi)=0$ implies that $\xi=\Gamma(F)$ for all $F\in\mathcal{F}$ and for all $\xi\in\Xi$, it is a *strict $\mathcal{F}$-identification function* for Γ. A functional $\Gamma\colon \mathcal{F}\to\Xi$ is called *identifiable* if there is a strict $\mathcal{F}$-identification function for it.

In forecast performance evaluation, (strict) identification functions are deployed to check calibration of forecasts (Nolde and Ziegel, 2017; Dimitriadis et al., 2019), akin to a goodness-of-fit test. That means they assess the absolute performance of a forecast method rather than the relative performance with respect to another method.

Following classical terminology we distinguish three cases in the choice of the dimensionality ℓ of φ: (i) the *unidentified case* $\ell<k$; (ii) the *exactly identified case* $\ell=k$; and (iii) the *overidentified case* $\ell>k$. In this paper, we shall restrict attention to the exactly identified case where the dimension of the action domain Ξ coincides with the dimension of the identification function. The reasons are two-fold (compare also to Remark 3.3): First, exploiting first-order conditions, the gradient of a consistent loss function constitutes an identification function whose dimension coincides with the dimension of the action domain; see Example 2.11 and Subsection 2.4. The second one is a novel characterisation result for the class of strict $\mathcal{F}$-identification functions for a functional Γ. Under certain richness assumptions on $\mathcal{F}$ (see Theorem B.1 for a precise formulation) and if $\varphi\colon \mathbb{R}\times\Xi\to\mathbb{R}^k$ is a strict $\mathcal{F}$-identification function, the entire class of strict $\mathcal{F}$-identification functions is given as

$$\big\{h(\xi)\varphi(y,\xi)\,|\,h\colon \Xi\to\mathbb{R}^{k\times k},\ \det(h(\xi))\neq 0 \text{ for all } \xi\in\Xi\big\}. \tag{2.5}$$

As a matter of fact we are aware of such a characterisation result only for the exactly identified case $\ell=k$; see Remark B.2 for a further discussion.

Recalling the argument of Remark 2.1 that Assumption (1) induces a functional $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta$, we can modify Definition 2.7 to account for this situation. For estimation of semiparametric models, and similar to Definition 2.4, we introduce an unconditional and a conditional notion, where the latter coincides with the notion of conditional moment conditions (or restrictions), as given, e.g., in Newey (1993) and the references therein.

Definition 2.8. Under Assumption (1), let $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta \subseteq \mathbb{R}^q$ be the functional given by (2.1). Let $\psi\colon \mathbb{R} \times \mathbb{R}^p \times \Theta \to \mathbb{R}^\ell$ be a function such that $\mathbb{E}\|\psi(Y, X, \theta)\|_1 < \infty$ for all $(Y, X) \in \mathcal{Z}$ and for all $\theta \in \Theta$.

(i) The function ψ is an *unconditional $\mathcal{F}_\mathcal{Z}$-identification function* for $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta$ if

$$\mathbb{E}\big[\psi(Y, X, \theta_0(F_Z))\big] = 0 \qquad \text{for all } Z = (Y, X) \in \mathcal{Z}\,.$$

It is a *strict unconditional $\mathcal{F}_\mathcal{Z}$-identification function* for θ_0 if additionally

$$\Big(\mathbb{E}[\psi(Y, X, \theta)] = 0 \implies \theta = \theta_0(F_Z)\Big) \qquad \text{for all } Z = (Y, X) \in \mathcal{Z},\ \theta \in \Theta\,.$$

(ii) The function ψ is a *conditional $\mathcal{F}_\mathcal{Z}$-identification function* for $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta$ if

$$\mathbb{E}\big[\psi(Y, X, \theta_0(F_Z))\big|X\big] = 0 \quad \text{a.s.} \qquad \text{for all } Z = (Y, X) \in \mathcal{Z}\,.$$

It is a *strict conditional $\mathcal{F}_\mathcal{Z}$-identification function* for θ_0 if additionally

$$\Big(\mathbb{E}\big[\psi(Y, X, \theta)\big|X\big] = 0 \quad \text{a.s.} \implies \theta = \theta_0(F_Z)\Big) \qquad \text{for all } Z = (Y, X) \in \mathcal{Z},\ \theta \in \Theta\,.$$

Similarly to Remark 2.6, a (strict) unconditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0 can be considered as a (strict) identification function in the sense of Definition 2.7 with observation domain $\mathbb{R} \times \mathbb{R}^p$ and action domain Θ. The following lemma establishes the counterparts of Theorem 2.5 (i) and (ii), with similar attenuations with respect to the *strictness* as in Theorem 2.5 (ii).

Lemma 2.9. *Under Assumption* (1) *the following implications hold for a map* $\varphi\colon \mathbb{R} \times \Xi \to \mathbb{R}^\ell$*:*

(i) *If* φ *is a (strict)* $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$*-identification function for* Γ *then* $\mathbb{R} \times \mathbb{R}^p \times \Theta \ni (y, x, \theta) \mapsto \varphi(y, m(x, \theta))$ *is a (strict) conditional* $\mathcal{F}_\mathcal{Z}$*-identification function for* θ_0*.*

(ii) *If* $(y, x, \theta) \mapsto \varphi(y, m(x, \theta))$ *is a conditional* $\mathcal{F}_\mathcal{Z}$*-identification function for* θ_0 *then* φ *is a* $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$*-identification function for* Γ*.*

To arrive at a counterpart of Theorem 2.5 (iii) note that, by the tower property, any (strict) conditional $\mathcal{F}_\mathcal{Z}$-identification for θ_0 is an unconditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0. However, it generally fails to be strict; see Example B.3. To construct a *strict* unconditional $\mathcal{F}_\mathcal{Z}$-identification function, recall the well-known fact that $\mathbb{E}\big[\varphi\big(Y, m(X, \theta)\big)\big|X\big] = 0$ almost surely is equivalent to

$$\mathbb{E}\big[a(X)^\intercal \varphi\big(Y, m(X, \theta)\big)\big] = 0 \qquad \text{for all measurable } a\colon \mathbb{R}^p \to \mathbb{R}^k, \tag{2.6}$$

such that $\mathbb{E}\big[\big|a(X)^\intercal \varphi\big(Y, m(X, \theta)\big)\big|\big] < \infty$ (or alternatively for all bounded for for all indicator functions). Since (2.6) is a pointwise statement for all $\theta \in \Theta$, the functions $a(X)^\intercal$ may additionally depend on θ, such that we write $a(X, \theta)^\intercal$. To render the equivalence in (2.6) statistically feasible, one needs to effectively reduce the number of functions $a(X,\theta)^\intercal$ to be finite, say s. Then, the classical approach is to stack them on top of each other, creating an *instrument matrix* $A(X,\theta) \in \mathbb{R}^{s\times\ell}$, see e.g. Newey (1990) and the references therein for the estimation context and Newey (1985), Bierens (1990), Nolde and Ziegel (2017) for similar procedures for testing conditional moment restrictions.

If there is some finite set $B \subseteq \mathbb{R}^p$ such that $\mathbb{P}(X \in B) = 1$ for all $X \in \mathcal{X}$ (e.g. if they are categorical), one can easily select finitely many instrument variables by employing appropriate indicator functions. For the general situation when the support of X is not finite, Newey and McFadden (1994, p. 2127) discuss that it is usually not feasible to find primitive conditions for strict unconditional identifiability and Domínguez and Lobato (2004) provide examples that an arbitrary choice of A may lead to a non-strict unconditional identification function. We refer to Rothenberg (1971), Brown (1983), Roehrig (1988) and Komunjer (2012) for some specific positive results. In the same direction, Proposition B.4 generalises the well-known condition from linear mean regression that $\mathbb{E}[XX^\intercal]$ must be of full rank. It is, thus, a sort of counterpart of Theorem 2.5 (iii). Nevertheless, in line with Newey and McFadden (1994, p. 2127), in most practical situations, one needs to show the strict unconditional identifiability on a case by case basis or one needs to simply assume it. For a partial counterpart of Theorem 2.5 (iv) we refer to Proposition B.5, and to Proposition C.1 for unique identification results specifically for double quantile models.

In the remainder of the paper, we make the following choices: Following Lemma 2.9, we solely work with strict conditional $\mathcal{F}_\mathcal{Z}$-identification functions of the form $\varphi\big(y, m(x,\theta)\big)$ where φ is a strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function for Γ. Following the discussion around (2.6), we restrict attention to $(q\times k)$-instrument matrices, such that we obtain unconditional identification functions

$$\psi_A(y,x,\theta) = A(x,\theta)\varphi\big(y, m(x,\theta)\big). \tag{2.7}$$

In light of the characterisation result of Theorem B.1, this has the advantage that the choice of the identification function φ in (2.7) is actually irrelevant. Indeed, suppose one considers $\varphi'\colon \mathbb{R}\times\Xi \to \mathbb{R}^k$ rather than φ in (2.7). This means there is a matrix-valued function $h\colon \operatorname{int}(\Xi) \to \mathbb{R}^{k\times k}$ of full rank such that $\varphi'\big(y, m(x,\theta)\big) = h\big(m(x,\theta)\big)\varphi\big(y, m(x,\theta)\big)$. Then we can use the matrix $A'(x,\theta) = A(x,\theta)\big(h(m(x,\theta))\big)^{-1}$ such that $A'(x,\theta)\varphi'\big(y, m(x,\theta)\big) = A(x,\theta)\varphi\big(y, m(x,\theta)\big)$. Consequently, it is no loss of generality to fix a certain strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function φ for Γ since the remaining flexibility can always be captured through the choice of the instrument matrix A.

2.4. Connection between loss and identification functions

As already noted in the Introduction, there is an intimate relationship between (strictly) consistent loss functions and strict identification functions for Γ via differentiation and integration. The intuition is that, under sufficient smoothness and regularity conditions, first-order conditions yield that the derivative of a (strictly) consistent loss for Γ is an identification function. While this leads to an (almost) one-to-one relation between consistent losses and identification functions for one-dimensional functionals, it turns out that there are considerably more identification functions than consistent losses if Γ is multivariate. This disparity proves to be consequential for efficient estimation of semiparametric models for vector-valued functionals, as discussed in the subsequent sections of this article. More details are in order:

If Γ is univariate, its mixture-continuity[3] implies that every strictly consistent loss ρ is

[3] Γ is mixture-continuous on a convex class $\mathcal{F}$ if for any $F, G \in \mathcal{F}$ $[0,1] \ni \lambda \mapsto \Gamma((1-\lambda)F + \lambda G)$ is continuous.

(strictly) *order-sensitive* meaning that $\xi \mapsto \bar{\rho}(F,\xi)$ is (strictly) decreasing (increasing) for $\xi \le \Gamma(F)$ (for $\xi \ge \Gamma(F)$); see Nau (1985, Proposition 3), Lambert (2013, Proposition 2) and Bellini and Bignozzi (2015, Proposition 3.4). Therefore, the derivative of ρ is an *oriented* identification function in the sense that $\nabla_\xi \bar{\rho}(F,\xi) \le 0$ (≥ 0) if $\xi \le \Gamma(F)$ ($\ge \Gamma(F)$). Intuitively, this excludes the existence of additional local minima of the expected loss, while possible saddle points still remain an issue. Moreover, Osband's principle in dimension one (Lambert et al., 2008; Gneiting, 2011a) implies that – under sufficient regularity conditions – if φ is an identification function for Γ, then for any consistent loss ρ there is a real-valued function h such that

$$\nabla_\xi \bar{\rho}(F,\xi) = h(\xi)\bar{\varphi}(F,\xi) \tag{2.8}$$

for all $\xi \in \Xi$ and $F \in \mathcal{F}$. Even if ρ is strictly consistent, its derivative is not necessarily a strict identification function due to possible saddle points of the expected loss. That means even if φ is strict, h might vanish at some points, see Steinwart et al. (2014) and Newey and McFadden (1994, p. 2117) for further details and examples. If φ is oriented and strict, then h is non-negative. This means, on the other hand, that we can also start with an oriented strict identification function, multiply it with any positive h, and integrate it. This results in a strictly order-sensitive, and therefore, strictly consistent loss (Steinwart et al., 2014, Theorem 7). If φ is not strict and h simply non-negative, the resulting loss is merely consistent. This leads to the fact that there is a one-to-one relation between consistent losses and *oriented* identification functions for Γ.[4]

Osband's principle is also available for multivariate functionals Γ mapping to $\Xi \subseteq \mathbb{R}^k$ with $k > 1$. Then (2.8) holds equivalently with h being a $(k \times k)$-matrix, and second-order conditions imply that the Hessian $\nabla^2 \bar{\rho}(F,\xi)$ must be positive semi-definite at $\xi = \Gamma(F)$. Moreover, under sufficient smoothness assumptions, the second order partial derivatives of $\bar{\rho}(F,\xi)$ must commute, implying that the Hessian must be symmetric for each $\xi \in \Xi$ and for each $F \in \mathcal{F}$. This, in turn, imposes strong conditions on h (Fissler and Ziegel, 2016, Corollary 3.3). Without these symmetry conditions, $h(\xi)\bar{\varphi}(F,\xi)$ even fails to be integrable (Königsberger, 2004, p. 185). For the construction of losses via integration this means that, starting with a (strict) identification function, there are severe limitations. While any h (with $\det(h(\xi)) \neq 0$) induces a (strict) identification function (see the discussion around (2.5) and Theorem B.1), by far not any such (strict) identification function is integrable, or even gives rise to a (strictly) consistent loss function, which can be illustrated with the following example.

Example 2.10. Consider the double quantile $\Gamma = (Q_\alpha, Q_\beta)$ with the strict identification function $\varphi(y,\xi_1,\xi_2) = \big(\mathbb{1}_{\{y \le \xi_1\}} - \alpha,\ \mathbb{1}_{\{y \le \xi_2\}} - \beta\big)^\intercal$. Fissler and Ziegel (2016, Proposition 4.2(i)) and Fissler and Ziegel (2019a) yield that the derivative of any expected (strictly) consistent loss function takes the form $h(\xi_1,\xi_2)\bar{\varphi}(F,\xi_1,\xi_2)$ where $h(\xi_1,\xi_2) = \operatorname{diag}(w_1(\xi_1), w_2(\xi_2))$ and the diagonal entries are non-negative. On the other hand, there is evidently a considerably larger class of functions h to $\mathbb{R}^{2\times 2}$ such that $\det(h(\xi_1,\xi_2)) \neq 0$ for all $(\xi_1,\xi_2) \in \Xi$.

It is precisely this fact where the gap between the class of (strictly) consistent losses and strict identification functions stems from for multivariate functionals. By virtue of the arguments in Sections 2.2 and 2.3, this gap in turn induces a gap between the classes of consistent M- and Z-estimators, which is illustrated by the following Example 2.11. Note that in order to use

[4] Note that if the identification function fails to be oriented, it can still be integrated, but does not yield a consistent score. E.g., $\varphi(y,\xi) = (\mathbb{1}\{\xi \ge 0\} - \mathbb{1}\{\xi < 0\})(\xi - y)$ is a strict identification function for the mean which fails to be oriented. It is easy to check that the integral $\rho(y,\xi) = \int_0^\xi \varphi(y,z)\,\mathrm{d}z$ is not a strictly consistent loss function for the mean. This identification function constitutes a counterexample to Steinwart et al. (2014, Lemma 6).

Osband's principle in Example 2.11, and in line with the discussion in Newey and McFadden (1994, Chapter 7), it is sufficient to assume that the conditional expectation $\mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big|X\big]$ is differentiable in θ almost surely for all $(Y, X) \in \mathcal{Z}$. This allows us to treat also losses that are *per se* not differentiable, such as the pinball loss.

Example 2.11. Let Assumption (1) hold for some k-dimensional functional $\Gamma\colon \mathcal{F}_{\mathcal{Y}|\mathcal{X}} \to \Xi$ with a strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function $\varphi\colon \mathbb{R} \times \Xi \to \mathbb{R}^k$ and a strictly $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss function $\rho\colon \mathbb{R} \times \Xi \to \mathbb{R}$. Suppose that $\mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big|X\big]$ is differentiable in θ almost surely for all $(Y, X) \in \mathcal{Z}$. Under the richness conditions on $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ of Fissler and Ziegel (2016, Theorem 3.2), we then have

$$\nabla_\theta \mathbb{E}\big[\rho\big(Y, m(X,\theta)\big)\big|X\big] = \nabla_\theta m(X,\theta)^\intercal \cdot h\big(m(X,\theta)\big) \cdot \mathbb{E}\big[\varphi\big(Y, m(X,\theta)\big)\big|X\big]\,, \tag{2.9}$$

where h takes values in $\mathbb{R}^{k\times k}$ and the gradient $\nabla_\theta m(X,\theta)$ is in $\mathbb{R}^{k\times q}$. Comparing (2.9) with (2.7), one obtains the identity $A(X,\theta) = \nabla_\theta m(X,\theta)^\intercal \cdot h\big(m(X,\theta)\big)$ for the instrument matrix. The presence of h and the limitations of the choice of h discussed above yield that there are considerably fewer (strict) model-consistent losses than (strict) moment functions.

3. Efficient Estimation

In this section, we consider implications of the gap between the classes of loss and identification functions established in the previous section on efficient (in the sense of Newey (1993)) M- and Z-estimation for semiparametric models for vector-valued functionals. For further implications of this gap on equivariant estimation, we refer to Appendix A. We extend the notation introduced in the beginning of Section 2.1 and henceforth denote the conditional distribution by $F_t := F_{Y_t|X_t}$ and whenever they exist, the corresponding conditional density by f_t, the conditional expectation by $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|X_t]$ and the conditional variance by $\mathrm{Var}_t(\cdot) := \mathrm{Var}(\cdot|X_t)$. Henceforth, we assume that the regularity conditions in Assumptions (1) – (4) hold.

Following the results of Theorem 2.5 and the discussion thereafter, employing strictly $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss functions for the functional under consideration is a necessary condition for consistent M-estimation of semiparametric models. Importantly, as indicated in (1.2) and (1.6), the loss and identification function may be time dependent. Following the discussion after (2.7), the time dependence of ψ_t in (1.6) can be captured entirely by time varying instrument matrices $A_t(X_t, \theta)$, leading to

$$\psi_t\big(Y_t, X_t, \theta\big) = \psi_{A_t}\big(Y_t, X_t, \theta\big) = A_t(X_t, \theta)\varphi\big(Y_t, m(X_t, \theta)\big). \tag{3.1}$$

For notational convenience, let $\mathbb{A}$ denote the sequence of instrument matrices $(A_t)_{t\in\mathbb{N}}$, and $\widehat{\theta}_{Z,T,\mathbb{A}}$ the Z-estimator at (1.6) based on $\mathbb{A}$ at (3.1).

Henceforth, we assume that the considered M- and Z-estimators are consistent and asymptotically normal. Primitive conditions for this (in the time-series context) are widely available, see e.g. Huber (1967), Weiss (1991), Newey and McFadden (1994), Andrews (1994), Davidson (1994) and van der Vaart (1998). These conditions include classical moment and dependence conditions on the process $(Y_t, X_t)_{t\in\mathbb{N}}$ together with smoothness assumptions on the (conditional expectations of the) employed loss and identification functions, and an identification condition for the model parameters. While for the M-estimator, this identification condition is fulfilled through Theorem 2.5 by employing strictly consistent loss functions, the analogue condition that ψ_t are strict $\mathcal{F}_{\mathcal{Z}}$-identification functions for θ_0 for all $t \in \mathbb{N}$ has to be imposed separately and be usually verified on a case-by-case basis; see Section 2.4 and Appendix B for further details and Appendix C for specific results for the double quantile model.

Given that such conditions hold for the Z-estimator $\widehat{\theta}_{Z,T,\mathbb{A}}$, and further given that the sequence $\psi_t\big(Y_t, X_t, \theta_0\big)$ is uncorrelated,[5] it holds that

$$\Sigma_{T,\mathbb{A}}^{-1/2} \Delta_{T,\mathbb{A}} \sqrt{T} \big(\widehat{\theta}_{Z,T,\mathbb{A}} - \theta_0\big) \xrightarrow{d} \mathcal{N}(0, I_q), \tag{3.2}$$

where the asymptotic covariance is governed by the terms

$$\Sigma_{T,\mathbb{A}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[A_t(X_t, \theta_0) S_t(X_t, \theta_0) A_t(X_t, \theta_0)^\intercal\right] \in \mathbb{R}^{q\times q} \qquad \text{and} \tag{3.3}$$

$$\Delta_{T,\mathbb{A}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[A_t(X_t, \theta_0) D_t(X_t, \theta_0)\right] \in \mathbb{R}^{q\times q}, \tag{3.4}$$

where, for any $\theta \in \Theta$,

$$S_t(X_t, \theta) = \mathbb{E}_t\left[\varphi\big(Y_t, m(X_t, \theta)\big)\varphi\big(Y_t, m(X_t, \theta)\big)^\intercal\right] \in \mathbb{R}^{k\times k} \qquad \text{and} \tag{3.5}$$

$$D_t(X_t, \theta) = \nabla_\theta \mathbb{E}_t\big[\varphi\big(Y_t, m(X_t, \theta)\big)\big]^\intercal \in \mathbb{R}^{k\times q}. \tag{3.6}$$

Generally speaking, the asymptotic distribution of an M-estimator corresponds to the asymptotic distribution of the Z-estimator whose identification function of the form (3.1) is matched by the derivative (with respect to θ) of the underlying loss function of the M-estimator, see e.g. Theorems 3.1, 3.2, and the discussion on p. 2145 in Newey and McFadden (1994), and Example 2.11 of this article. If the loss is not differentiable, but the conditional expectation is almost surely, then this rationale also holds on the level of the conditional expectations (Newey and McFadden, 1994, Theorems 7.1 und 7.2). Consequently, in the sequel we say that an M-estimator has an equivalent Z-estimator if the derivative of (the conditional expectation of) the loss function equals (the conditional expectation of) the identification function almost surely.[6] Also, M-estimators based on $(\rho_t)_{t\in\mathbb{N}}$ and $(\tilde{\rho}_t)_{t\in\mathbb{N}}$ have the same asymptotic covariance if there is some $c > 0$ and functions $\kappa_t \colon \mathbb{R} \to \mathbb{R}$ such that $\tilde{\rho}_t\big(Y_t, m(X_t, \theta)\big) = c\rho_t\big(Y_t, m(X_t, \theta)\big) + \kappa_t(Y_t)$. This is why we dispense with a discussion of the terms c and κ_t in the sequel.

We say that an asymptotically normal estimator is *efficient* if there is no other asymptotically normal estimator with a smaller covariance matrix in the *Loewner order*, where for two positive semi-definite matrices A and B, we say that $A \succcurlyeq B$ if and only if $A - B$ is positive semi-definite, and $A \succ B$ if and only if $A - B$ is positive definite. Motivated by the discussion in (Newey, 1990, p. 102), we deliberately omit a discussion of superefficient estimators in this article. The following theorem establishes necessary and sufficient conditions for efficient Z-estimation, by restating and extending the theory of Hansen (1985), Chamberlain (1987), and Newey (1993).

Theorem 3.1. *Under Assumptions (1) –(4), let φ be a strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function for Γ. Let $\widehat{\theta}_{Z,T,\mathbb{A}^*}$ be the Z-estimator given at (1.6) which is asymptotically normal and which is based on the strict unconditional $\mathcal{F}_{\mathcal{Z}}$-identification function at (3.1) for the instrument matrices*

[5] This condition implies the form of $\Sigma_{T,\mathbb{A}}$ given in (3.3), i.e. the "HAC terms" in the general form vanish. The uncorrelatedness condition is e.g. satisfied in the standard time series case if X_t contains finitely many lags of Y_t and if $\psi_t\big(Y_t, X_t, \theta_0\big)$ is a martingale difference sequence.

[6] This equivalence is understood in terms of their first-order asymptotics and we do not consider the case of higher-order asymptotics in the sense of Newey and Smith (2004), where these estimators may of course behave differently.

$A^*_{t,C}(X_t, \theta)$, which are such that

$$A^*_{t,C}(X_t, \theta_0) = C D_t(X_t, \theta_0)^\intercal S_t(X_t, \theta_0)^{-1} \quad \text{for all } t \in \mathbb{N}, \tag{3.7}$$

where $S_t(X_t, \theta_0)$ and $D_t(X_t, \theta_0)$ are given at (3.5) and (3.6), assuming that $S_t(X_t, \theta_0)$ is invertible, and C is any deterministic and invertible $q \times q$ matrix. Then:

(i) *The Z-estimator $\widehat{\theta}_{Z,T,\mathbb{A}^*}$ has the asymptotic covariance matrix*

$$\Lambda_T^{-1} := \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[D_t(X_t, \theta_0)^\intercal S_t(X_t, \theta_0)^{-1} D_t(X_t, \theta_0) \right] \right)^{-1}. \tag{3.8}$$

(ii) *For any sequence of instrument matrices $\mathbb{A} = (A_t)_{t \in \mathbb{N}}$, and $\Delta_{T,\mathbb{A}}$, $\Sigma_{T,\mathbb{A}}$ as given at (3.3) and (3.4), it holds that $\Delta^{-1}_{T,\mathbb{A}} \Sigma_{T,\mathbb{A}} \Delta^{-1}_{T,\mathbb{A}} \succcurlyeq \Lambda_T^{-1}$ for all $T \geq 1$.*

(iii) *If for some $t \in \{1, \ldots, T\}$ and for any non-singular and deterministic matrix C it holds that $\mathbb{P}\big(A_t(X_t, \theta_0) \neq A^*_{t,C}(X_t, \theta_0)\big) > 0$, then $\Delta^{-1}_{T,\mathbb{A}} \Sigma_{T,\mathbb{A}} \Delta^{-1}_{T,\mathbb{A}} \succ \Lambda_T^{-1}$.*

Parts (i) and (ii) of Theorem 3.1 are direct time series generalizations of the efficiency result of Hansen (1985), Chamberlain (1987), and Newey (1993). Together, they state that Λ_T^{-1} is an asymptotic efficiency bound for the general Z-estimator for semiparametric models and that the Z-estimator based on the choice $A^*_{t,C}(X_t, \theta)$ for all $t \in \mathbb{N}$ which fulfills (3.7) attains this efficiency bound, and is consequently an efficient Z-estimator. Thus, parts (i) and (ii) of Theorem 3.1 can be understood as a sufficient condition for efficient semiparametric Z-estimation.

Conversely, part (iii) can be interpreted as a necessary condition for efficient estimation and is novel to the literature. It states that efficient semiparametric estimation can only be carried out by choosing instrument matrices satisfying (3.7) almost surely. This necessary condition for efficient estimation is crucial for the following sections where we show that for certain functionals, the M-estimator of semiparametric models cannot attain the Z-estimation efficiency bound and consequently neither the semiparametric efficiency bound in the sense of Stein (1956), which is further discussed in Remark 3.4.

Remark 3.2. Note that the efficient instrument matrix $A^*_{t,C}(X_t, \theta_0)$ depends on the choice of the conditional identification function φ. Thus, the exact form of the necessary and sufficient conditions on the instrument matrix $A_t(X_t, \theta_0)$ in Theorem 3.1 depend on the choice of φ. However, for two different choices of conditional identification functions, the respective efficient instrument matrices can be converted in a straight-forward fashion through the formulas after (2.7) (given that the class $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ satisfies the assumptions of Theorem B.1). Hence, the efficiency bound Λ_T^{-1} is invariant to the inessential choice of φ, and Theorem 3.1 can be interpreted as global, but φ-dependent, necessary and sufficient conditions for efficiency.

Remark 3.3. While overidentified GMM-estimation as defined in (1.4) with $s > q$ can generally improve efficiency compared to Z-estimation (see e.g. Hansen (1982), Newey and McFadden (1994), Hall (2005), among many others), when employing the efficient instrument choice in (3.7), there is no additional efficiency gain through using overidentifying moment restrictions. For details, see e.g. Newey (1993), and notice that the proofs of Theorem 3.1, (i) and (ii) work identically when including overidentifying moment restrictions together with a weighting matrix P_T as in (1.4). Consequently, without loss of generality, we restrict attention to efficient instrument Z-estimation in Theorem 3.1.

Remark 3.4 (Connections to the semiparametric efficiency bound)**.** We would like to stress that the main focus of this article lies on the Z-estimation efficiency bound for conditional moment restrictions of Hansen (1985), Chamberlain (1987) and Newey (1993). In the context of i.i.d. processes and differentiable moment conditions, Chamberlain (1987) shows that this bound coincides with the general semiparametric efficiency bound in the sense of a least favorable submodel of Stein (1956); c.f. Newey (1990) and Bickel et al. (1998) for surveys on this matter. For recent progress on semiparametric efficiency bounds and efficient semiparametric estimation, see e.g. Ackerberg et al. (2014), Ai and Chen (2003), Janková and van de Geer (2018), Hristache and Patilea (2016), and Komunjer and Vuong (2010b,a), among many others.

The definition of the semiparametric efficiency bound builds on the idea that the data stems from a *parametric submodel*, i.e., a parametric model which completely specifies the full distribution, contains the correctly specified model, and satisfies the semiparametric model assumption. E.g., if we consider a semiparametric model for the conditional mean, we do not make any assumptions about the exact conditional distribution beyond the mean assumption. Any model which parametrises the full conditional distribution (e.g., a normal distribution with parameterised variance) is such a *parametric submodel*. Estimation of any parametric submodel is subject to the classical Cramér-Rao efficiency bound, which can be attained, e.g., by maximum likelihood estimation using the true parametric distribution, dispensing with a discussion of superefficient estimators. For any parametric submodel, a consistent and asymptotically normal semiparametric estimator is contained in the class of estimators for this parametric submodel and thus, it is subject to the parametric Cramér-Rao efficiency bound. Consequently, any semiparametric estimator has an asymptotic variance which is no smaller than the Cramér-Rao bound for *any* parametric submodel. Hence, the *semiparametric efficiency bound* is defined as the supremum of the Cramér-Rao bounds of all parametric submodels.

The results of this paper concerning efficient estimation are derived with respect to the Z-estimation efficiency bound of Hansen (1985), Chamberlain (1987), and Newey (1993). In applications to smooth objective functions and i.i.d. processes, the result of Chamberlain (1987) can be used to equate these two bounds. However, as we are not aware of a general relation of these bounds for non-i.i.d. processes, we cannot preclude that the semiparametric efficiency bound is strictly smaller (in the Loewner order) than the Z-estimation efficiency bound in certain situations. Consequently, all following assertions are stated in relation to the Z-estimation efficiency bound. Notice that this does not affect our main conclusion in terms of efficient estimation: When the M-estimator cannot attain the Z-estimation efficiency bound, it also cannot attain the semiparametric efficiency bound, irrespective of whether these quantities coincide.

4. Semiparametric Models for Vector-Valued Functionals

We answer the question whether, and under which conditions, M-estimation can attain the Z-estimation efficiency bound in the sense of Newey (1993) (see Remark 3.4) for the pairs consisting of the first and second moment and (mean, variance) in Subsection 4.1, the double quantile model Subsection 4.2, and the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ in Subsection 4.3.

4.1. Semiparametric Models for Multiple Moments

We consider joint semiparametric models for the first and second moments, denoted by Γ_{mom}, and closely related, joint models for mean and variance, $\Gamma_{(\mathbb{E},\mathrm{Var})}$. Since mean and variance are considered as the most important functionals in classical statistics, the related class of ARMA-GARCH models (Bollerslev, 1986) is omnipresent in the econometric literature, and is often estimated through M- or Z-estimation. See also Spady and Stouli (2018) for joint regression models for the mean and the variance.

We assume that $\mathcal{F}_{\mathcal{Y}|\mathcal{X}} \subseteq \mathcal{F}_2$, where $\mathcal{F}_2$ is the class of distributions with finite second moment. Recalling that Γ_{mom} and $\Gamma_{(\mathbb{E},\text{Var})}$ are in bijection, we can invoke the *revelation principle* (Osband, 1985; Gneiting, 2011a; Fissler, 2017) to relate the corresponding strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification and strictly $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss functions. Strict identification functions are given by

$$\varphi_{\text{mom}}(y,\xi_1,\xi_2) = \begin{pmatrix} \xi_1 - y \\ \xi_2 - y^2 \end{pmatrix}, \qquad \text{and} \qquad \varphi_{(\mathbb{E},\text{Var})}(y,\xi_1,\xi_2) = \begin{pmatrix} \xi_1 - y \\ \xi_2 + \xi_1^2 - y^2 \end{pmatrix}. \tag{4.1}$$

Theorem 2.5 yields that the full class of consistent M-estimators at (1.2) is determined by the full class of (strictly) $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss functions. Using the revelation principle and following Fissler and Ziegel (2016, Proposition 4.4) and Fissler and Ziegel (2019a), if $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ is large enough (e.g., if it equals $\mathcal{F}_2$), the class of all differentiable (strictly) $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss functions is given by

$$\begin{aligned} \rho_{\text{mom},t}(y,\xi_1,\xi_2) &= -\phi_t(\xi_1,\xi_2) + \nabla\phi_t(\xi_1,\xi_2)^\intercal \begin{pmatrix} \xi_1 - y \\ \xi_2 - y^2 \end{pmatrix} + \kappa_t(y), \\ \rho_{(\mathbb{E},\text{Var}),t}(y,\xi_1,\xi_2) &= -\phi_t(\xi_1,\xi_2+\xi_1^2) + \nabla\phi_t(\xi_1,\xi_2+\xi_1^2)^\intercal \begin{pmatrix} \xi_1 - y \\ \xi_2 + \xi_1^2 - y^2 \end{pmatrix} + \kappa_t(y) \end{aligned} \tag{4.2}$$

where $\phi_t : \{(\xi_1,\xi_2) \in \mathbb{R}^2 \,|\, \xi_1^2 \le \xi_2\} \to \mathbb{R}$ are (strictly) convex and twice differentiable functions with gradient $\nabla\phi_t$, and κ_t is an $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-integrable function. For any sequence $\Phi = (\phi_t)_{t\in\mathbb{N}}$ of such functions, we denote the corresponding M-estimators defined via (1.2) by $\widehat{\theta}^{\text{mom}}_{M,T,\Phi}$ and $\widehat{\theta}^{(\mathbb{E},\text{Var})}_{M,T,\Phi}$.

Proposition 4.1. *Under Assumption (2), suppose that the M-estimators $\widehat{\theta}^{mom}_{M,T,\Phi}$ and $\widehat{\theta}^{(\mathbb{E},\text{Var})}_{M,T,\Phi}$ for the first two moments and for $(\mathbb{E},\text{Var})$ are asymptotically normal. If almost surely*

$$\phi_t(z) = \frac{1}{2} z^\intercal \mathbf{Var}_t\left(\left(Y_t, Y_t^2\right)\right)^{-1} z \qquad \text{for all } t \in \mathbb{N}, \tag{4.3}$$

then these M-estimators attain the corresponding Z-estimation efficiency bounds in (3.8).

This result is in line with the classical univariate mean regression, where both, M- and Z-estimators are able to attain the Z-estimation efficiency bound and the most efficient Bregman loss is given by the squared loss, weighted with the inverse of the conditional variance. Intuitively, this attainability can be explained by the fact that the classes of strictly consistent joint loss functions given in (4.2) are *relatively large* due to the presence of the general convex function ϕ_t, being a function in two arguments.

For the first two moments, this can be illustrated by comparing it to a minimal subclass in this context, namely the class only consisting of the sum of (strictly) consistent loss functions for the individual components, the first and second moment. This arises from (4.2) where ϕ_t takes the additive form $\phi_{\text{add},t}(\xi_1,\xi_2) = \phi_{1,t}(\xi_1) + \phi_{2,t}(\xi_2)$, where $\phi_{i,t}$ are both (strictly) convex. Since the Hessian $\nabla^2\phi_{\text{add},t}$ is diagonal, $\phi_{\text{add},t}$ can only take the form in (4.3) for the special situation when Y_t and Y_t^2 are conditionally uncorrelated. Since the class of convex functions on $\mathbb{R}^2$ is far larger than the sum of two convex functions in the individual components, the efficiency bound can be attained. For the pair of mean and variance, note that one cannot decompose the loss into a sum of strictly consistent losses for each component, due to the variance failing to be elicitable in general. In particular, this also shows the importance of modelling the variance *jointly* with the mean. However, an additive decomposition of ϕ_t as discussed above is also possible for mean and variance.

These results are in stark contrast to the double quantile (DQ) regression framework which we consider in Section 4.2, where the gap arises since the class of strictly consistent losses is relatively small, coinciding with the described minimal class.

4.2. Semiparametric Double Quantile Models

We consider semiparametric models for two quantiles at different levels, which arises naturally in the following fields of applications: In quantitative risk management, one is often interested in the VaR (the quantiles) of financial returns at two small probability levels, say 1% and 2.5%, which directly motivates modelling two (multiple) quantiles jointly through double quantile models; see e.g. Engle and Manganelli (2004), Koenker and Xiao (2006), Gourieroux and Jasiak (2008), White et al. (2015), Schmidt and Zhu (2016), Couperier and Leymarie (2019), Catania and Luati (2019). Furthermore, prediction intervals can naturally be defined as the interval spanned by two (conditional) quantiles (with levels of e.g. 5% and 95%); see Gneiting and Raftery (2007), Brehmer and Gneiting (2020), Fissler et al. (2020) for loss and identification functions for quantile based prediction intervals, and Bracher et al. (2020), Petropoulos and Makridakis (2020) and UMass-Amherst Influenza Forecasting Center of Excellence (2020) for timely applications in epidemiology covering, among others, the prediction of COVID-19 infection rates. Eventually, the entire conditional distribution can conveniently be approximated through multiple conditional quantiles, see e.g. Buchinsky (1994), Angrist et al. (2006), Chernozhukov et al. (2010), Schmidt and Zhu (2016) for microeconomics applications to earning distributions and unemployment insurance benefits and Adrian et al. (2019) for a macroeconomic application to uncertainty of GDP growth.

While in theory, semiparametric models for individual quantile levels can be estimated individually, one methodological demand on reasonable models for multiple quantiles is to impede quantile crossings (Koenker, 2005). Inter alia, this can be achieved either through certain parametric model structures or through parametric inequality restrictions; see e.g. Gourieroux and Jasiak (2008), White et al. (2015), Schmidt and Zhu (2016) and Catania and Luati (2019) for recently developed parametric model families for this.

Formally, we consider joint models for two different quantile levels $\alpha, \beta \in (0,1)$ where w.l.o.g. $\alpha < \beta$, given by $m(X_t, \theta) = \big(q_\alpha(X_t,\theta), q_\beta(X_t,\theta)\big)^\intercal$ for some parameter vector $\theta \in \Theta \subseteq \mathbb{R}^q$. While the results of this section can be extended to multiple quantiles at different levels in a straight-forward fashion, we restrict our attention to two quantile levels for a parsimonious exhibition. Let $\widehat{\theta}_{Z,T,\mathbb{A}}$ be the Z-estimator defined via (1.6) and (3.1) based on some sequence of instrument matrices $\mathbb{A}$ and the strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function

$$\varphi(y, \xi_1, \xi_2) = \begin{pmatrix} \mathbb{1}_{\{y \le \xi_1\}} - \alpha \\ \mathbb{1}_{\{y \le \xi_2\}} - \beta \end{pmatrix}, \tag{4.4}$$

assuming that all $F \in \mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ are differentiable at their correct α and β-quantiles with strictly positive derivatives. Following Theorem 3.1, efficient Z-estimation is based on the efficient instrument matrix $A^*_{t,C}(X_t,\theta_0) = C D_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$, where C is some deterministic and nonsingular matrix and where almost surely,

$$S_t(X_t,\theta_0) = \begin{pmatrix} \alpha(1-\alpha) & \alpha(1-\beta) \\ \alpha(1-\beta) & \beta(1-\beta) \end{pmatrix}, \qquad D_t(X_t,\theta_0) = \begin{pmatrix} f_t(q_\alpha(X,\theta_0))\nabla_\theta q_\alpha(X_t,\theta_0)^\intercal \\ f_t(q_\beta(X,\theta_0))\nabla_\theta q_\beta(X_t,\theta_0)^\intercal \end{pmatrix}. \tag{4.5}$$

Furthermore, Theorem 2.5 yields that the full class of consistent M-estimators at (1.2) is essentially determined by the full class of (strictly) $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent loss functions for the two quantiles. Fissler and Ziegel (2016, Proposition 4.2), (see also Fissler and Ziegel (2019a)) characterises this class, subject to smoothness conditions and richness assumptions on $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$, by

$$\begin{aligned}\rho_t(y,\xi_1,\xi_2) =& \big(\mathbb{1}_{\{y\le\xi_1\}}-\alpha\big)g_{1,t}(\xi_1)-\mathbb{1}_{\{y\le\xi_1\}}g_{1,t}(y)\\ &+\big(\mathbb{1}_{\{y\le\xi_2\}}-\beta\big)g_{2,t}(\xi_2)-\mathbb{1}_{\{y\le\xi_2\}}g_{2,t}(y)+\kappa_t(y),\end{aligned} \tag{4.6}$$

where κ_t is $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-integrable and $g_{1,t}:\mathbb{R}\to\mathbb{R}$ and $g_{2,t}:\mathbb{R}\to\mathbb{R}$ are (strictly) increasing for all $t\in\mathbb{N}$. Strikingly, this means that the whole class of (strictly) consistent losses for the double quantile coincides with the sum of (strictly) consistent losses for the individual quantiles. For any sequence $G=(g_{1,t},g_{2,t})_{t\in\mathbb{N}}$ of such functions, we denote the corresponding M-estimators defined via (1.2) by $\widehat{\theta}_{M,T,G}$.

In the following, we assume that both estimators are consistent and asymptotically normal. Most primitive conditions for this are straight-forward but tedious to verify, see e.g. Huber (1967), Newey and McFadden (1994), Andrews (1994), and specifically for semiparametric quantile models, Koenker and Bassett (1978), Giacomini and Komunjer (2005), Komunjer (2005), and Komunjer and Vuong (2010b). These conditions mainly consist of an absolutely continuous conditional distribution $F_t=F_{Y_t|X_t}$ with a strictly positive density f_t at the considered quantile(s), the existence of certain moments and of a controlled serial dependence structure of the underlying process. In addition, one has to verify strict unconditional model consistency of $\rho_t\big(Y_t,m(X_t,\theta)\big)$ for the M-estimator, and strict unconditional identification for the Z-estimator. Under Assumption (1), the former is guaranteed by Theorem 2.5 (i) and (iii) when using the strictly consistent losses at (4.6). For the latter, we refer to Proposition C.1 in Appendix C which shows strict identification for the efficient Z-estimator and for linear models. For general results, see Lemma 2.9 (i) and Proposition B.5.

The following theorem establishes that, under certain conditions, the M-estimator of the double quantile model is subject to the *efficiency gap*, i.e., it cannot attain the Z-estimation efficiency bound, and consequently neither the semiparametric efficiency bound.

Theorem 4.2. *Suppose that Assumptions (1) – (4) hold, $\widehat{\theta}_{M,T,G}$ is asymptotically normal and the following further regularity conditions hold:*

(DQ1) *The parameters of the individual models are separated, i.e., the model is of the form $m(X_t,\theta)=\big(q_\alpha(X_t,\theta^\alpha),q_\beta(X_t,\theta^\beta)\big)^\intercal$, where $\theta=\big(\theta^\alpha,\theta^\beta\big)\in\operatorname{int}\Theta\subseteq\mathbb{R}^q$, with $\theta^\alpha\in\mathbb{R}^{q_1}$ and $\theta^\beta\in\mathbb{R}^{q_2}$ and $q_1+q_2=q$.*

(DQ2) *For all $t\in\mathbb{N}$, and for all $A\in\mathcal{A}$ with $\mathbb{P}(A)=1$ there are q_1+1 mutually different $v_1,\dots,v_{q_1+1}\in\big\{\nabla_{\theta^\alpha}q_\alpha(X_t(\omega),\theta_0^\alpha)\in\mathbb{R}^{q_1}:\omega\in A\big\}\subseteq\mathbb{R}^{q_1}$, such that any subset of cardinality q_1 of $\{v_1,\dots,v_{q_1+1}\}$ is linearly independent. The analogue assertion holds for the gradient $\nabla_{\theta^\beta}q_\beta(X_t,\theta_0^\beta)$, replacing q_1 by q_2.*

(DQ3) *For all $t\in\mathbb{N}$, F_t is differentiable at $q_\alpha(X_t,\theta_0^\alpha)$ and $q_\beta(X_t,\theta_0^\beta)$ where the derivatives satisfy $f_t(q_\alpha(X_t,\theta_0^\alpha))>0$ and $f_t(q_\beta(X_t,\theta_0^\beta))>0$ almost surely, and $g'_{1,t}(\xi_1)>0$, $g'_{2,t}(\xi_2)>0$ for all ξ_1,ξ_2.*

Then, the following statements hold:

(A) *Suppose that for all $t\in\mathbb{N}$, $\nabla_{\theta^\alpha}q_\alpha(X_t,\theta_0^\alpha)=\nabla_{\theta^\beta}q_\beta(X_t,\theta_0^\beta)$ almost surely. The M-estimator $\widehat{\theta}_{M,T,G}$ attains the Z-estimation efficiency bound in (3.8) if and only if the following three* *conditions hold:*

$$\exists c_1 > 0\ \forall t \in \mathbb{N}: \quad f_t(q_\alpha(X_t, \theta_0^\alpha)) = c_1 f_t(q_\beta(X_t, \theta_0^\beta)) \quad a.s., \tag{4.7}$$

$$\exists c_2 > 0\ \forall t \in \mathbb{N}: \quad g_{1,t}'(q_\alpha(X_t, \theta_0^\alpha)) = c_2 f_t(q_\alpha(X_t, \theta_0^\alpha)) \quad a.s., \tag{4.8}$$

$$\exists c_3 > 0\ \forall t \in \mathbb{N}: \quad g_{2,t}'(q_\beta(X_t, \theta_0^\beta)) = c_3 f_t(q_\beta(X_t, \theta_0^\beta)) \quad a.s. \tag{4.9}$$

(B) *Furthermore, if* (4.8) *or* (4.9) *is violated, then* $\widehat{\theta}_{M,T,G}$ *does not attain the Z-estimation efficiency bound in* (3.8).

A discussion of the conditions of Theorem 4.2 is in order. The separated parameter condition (DQ1) contains a large class of possible models. Importantly, this does not imply that the parameter space is necessarily a Cartesian product of individual parameter spaces Θ^α and Θ^β, which can be relevant e.g. in order to impede quantile crossings through inequality restrictions on the parameter space. Condition (DQ2) concerns the variability of the model gradient, which is slightly stronger than the classical assumption on one-dimensional models m that the matrix $\mathbb{E}\left[\nabla_\theta m(X_t, \theta_0) \nabla_\theta m(X_t, \theta_0)^\intercal\right]$ be of full rank for all $t \in \mathbb{N}$.[7] Condition (DQ3) is standard for asymptotic normality of quantile regression parameters.

The additional gradient condition $\nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) = \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta)$ in (A) is mainly motivated through linear models, where these gradients are simply X_t. In contrast, statement (B) holds for general semiparametric models with separated parameters, but does not provide sufficient conditions for efficient M-estimation.

For the remainder of this subsection, we assume the gradient condition $\nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) = \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta)$, putting us in the situation of (A). Then, the core condition of this theorem on the underlying process is (4.7). Given that (4.7) holds, the remaining conditions (4.8) and (4.9) are fulfilled by using the obvious choices

$$g_{1,t}(\xi_1) = F_t(\xi_1), \qquad \text{and} \qquad g_{2,t}(\xi_2) = F_t(\xi_2), \tag{4.10}$$

for all $\xi_1, \xi_2 \in \mathbb{R}$ and for all $t \in \mathbb{N}$, given that F_t is strictly increasing and differentiable at the true α- and β-quantiles. These conditions coincide with classical efficient semiparametric quantile estimation (for one quantile level) in Komunjer and Vuong (2010b,a). We refer to (4.10) as the *pseudo-efficient* choices as they attain the Z-estimation efficiency bound only in certain situations.

In contrast to the semiparametric estimation of the first two moments, or of mean and variance, as discussed in the previous section, Theorem 4.2 demonstrates that an efficiency gap can arise for M-estimators for the double quantile model. The underlying reason is the relatively narrow class of strictly consistent loss functions given in (4.6), being simply the sum of strictly consistent losses for the individual quantiles. In contrast, the class of strict identification functions for this pair is considerably richer, see Example 2.10.

In the following, we analyze the validity of the core condition (4.7) in detail for the ubiquitous class of location-scale processes,

$$Y_t = \mu(X_t, \eta_0) + \sigma(X_t, \eta_0) u_t, \tag{4.11}$$

[7] E.g. consider a linear model with explanatory variable $X_t = (1, W_t)^\intercal$, where W_t attains only 0 and 1 with positive probability. Then, $\mathbb{E}\left[\nabla_\theta m(X_t, \theta_0) \nabla_\theta m(X_t, \theta_0)^\intercal\right] = \mathbb{E}\left[X_t X_t^\intercal\right]$ is positive definite whereas condition (DQ2) is not fulfilled. However, if W_t attains at least two different values with positive probability (or if its distribution is absolutely continuous), this condition holds.

where $\mu(X_t, \eta_0)$ and $\sigma(X_t, \eta_0)$ are parametric models for the conditional mean and scale of Y_t given X_t. The innovations $(u_t)_{t\in\mathbb{N}}$ are themselves independent, and independent of $(X_t)_{t\in\mathbb{N}}$, and we impose the semiparametric quantile stationarity condition that $z_\alpha = F_{u_t}^{-1}(\alpha)$ and $z_\beta = F_{u_t}^{-1}(\beta)$ are time-independent, such that Assumption (2) is satisfied. As long as this condition holds, we can allow for heterogeneously distributed innovations. Clearly, the conditional quantiles at level $\alpha \in (0,1)$ are given by $Q_\alpha(Y_t|X_t) = \mu(X_t, \eta_0) + \sigma(X_t, \eta_0) z_\alpha$.

The density transformation formula yields that $f_t\big(q_\alpha(X_t, \theta_0)\big) = f_{u_t}(z_\alpha)/\sigma(X_t, \eta_0)$. Consequently, condition (4.7) is equivalent to

$$\frac{f_t\big(q_\alpha(X_t,\theta_0)\big)}{f_t\big(q_\beta(X_t,\theta_0)\big)} = \frac{f_{u_t}(z_\alpha)}{f_{u_t}(z_\beta)} \overset{!}{=} c_1 \qquad \text{for all } t \in \mathbb{N}. \tag{4.12}$$

This implies that the M-estimator $\widehat{\theta}_{M,T,G}$ of the double quantile model for processes of the form (4.11) is able to attain the efficiency bound (based on the choices in (4.10)), if and only if the ratios of densities in (4.12) is constant in t. Consequently, for any i.i.d. innovations $(u_t)_{t\in\mathbb{N}}$, the M-estimator based on the choices (4.10) attains the Z-estimation efficiency bound.

However, one can easily construct examples where condition (4.12) is violated, e.g. by considering Student's t-distributed innovations $u_t \sim t_{\nu_t}(\mu_t, \sigma_t)$ with time-varying degrees of freedom ν_t, and where the time-varying means and standard deviations are given by

$$\mu_t = Q_\beta(t_{\nu_1}) - \sigma_t Q_\beta(t_{\nu_t}) \qquad \text{and} \qquad \sigma_t = \frac{Q_\alpha(t_{\nu_1}) - Q_\beta(t_{\nu_1})}{Q_\alpha(t_{\nu_t}) - Q_\beta(t_{\nu_t})}. \tag{4.13}$$

These choices are such that for $\alpha, \beta \in (0,1)$, $\alpha < \beta$, it holds that $Q_\alpha\left(t_{\nu_t}(\mu_t, \sigma_t)\right) = z_\alpha$ and $Q_\beta\left(t_{\nu_t}(\mu_t, \sigma_t)\right) = z_\beta$ for all $t \in \mathbb{N}$ and hence, the quantile-stationarity condition is satisfied while simultaneously condition (4.7) is violated for all quantile levels such that $\alpha \neq 1-\beta$.[8]

For centred or equal-tailed prediction intervals (Brehmer and Gneiting, 2020) with $\alpha = 1-\beta < 0.5$, we can choose skewed normally distributed (Azzalini, 1985) innovations $u_t \sim \mathcal{SN}(\mu_t, \sigma_t^2, \gamma_t)$ with time-varying skewness γ_t, where the means μ_t and the standard deviations σ_t are given by

$$\mu_t = Q_\beta\big(\mathcal{SN}(\gamma_1)\big) - \sigma_t Q_\beta\big(\mathcal{SN}(\gamma_t)\big), \qquad \sigma_t = \frac{Q_\alpha\big(\mathcal{SN}(\gamma_1)\big) - Q_\beta\big(\mathcal{SN}(\gamma_1)\big)}{Q_\alpha\big(\mathcal{SN}(\gamma_t)\big) - Q_\beta\big(\mathcal{SN}(\gamma_t)\big)}, \tag{4.14}$$

where $\mathcal{SN}(\gamma_1) := \mathcal{SN}(0,1,\gamma_1)$. Then, $Q_\alpha\left(\mathcal{SN}(\mu_t, \sigma_t^2, \gamma_t)\right) = z_\alpha$ and $Q_\beta\left(\mathcal{SN}(\mu_t, \sigma_t^2, \gamma_t)\right) = z_\beta$ for all $t \in \mathbb{N}$ and for all $\alpha, \beta \in (0,1)$, $\alpha < \beta$. We employ these models in the simulations for double quantile models in Appendix E, where we numerically confirm the theoretical claims of this section.

Constructing further processes where the M-estimator cannot attain the Z-estimation efficiency bound can be carried out along these lines, where the crucial condition is that these processes must go beyond the class of simple location-scale processes with i.i.d. residuals. Interesting candidates are the GAS-models of Creal et al. (2013), or for quantile-specific models, the CAViaR specification of Engle and Manganelli (2004).

4.3. Semiparametric Joint Quantile and ES Models

We consider a joint model for the quantile (or VaR) and ES at joint level $\alpha \in (0,1)$, given by $m(X_t, \theta) = \big(q_\alpha(X_t, \theta), e_\alpha(X_t, \theta)\big)^\intercal$, where $q_\alpha(X_t, \theta)$ is a model for the conditional α-quantile and

[8] Further notice that for all $t \in \mathbb{N}$, there is a deterministic $c_t > 0$ such that $f_t(q_\alpha(X_t, \theta_0^\alpha)) = c_t f_t(q_\beta(X_t, \theta_0^\beta)) > 0$ almost surely. Hence the efficient Z-estimator is strictly identified by Proposition C.1.

$e_\alpha(X_t, \theta)$ denotes a model for the ES at level α, ES_α. For a random variable Z with quantiles $Q_u(Z)$, the $\mathrm{ES}_\alpha(Z)$ is defined as $\frac{1}{\alpha}\int_0^\alpha Q_u(Z)\mathrm{d}u$, which simplifies to $\mathrm{ES}_\alpha(Z) = \frac{1}{\alpha}\mathbb{E}\left[Z\mathbb{1}_{\{Z\le q_\alpha(Z)\}}\right]$ if the distribution of Z is continuous at its α-quantile.

As shown by Gneiting (2011a) and Weber (2006), ES is generally neither elicitable nor identifiable and thus, applying Theorem 2.5 (ii) and (iv), Lemma 2.9 and Proposition B.5 provides evidence that both M- and Z-estimation of semiparametric models for the conditional ES standalone are generally infeasible. However, Fissler and Ziegel (2016, 2019a) show that under mild conditions, the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ is elicitable and identifiable, and further characterise the class of strictly consistent loss functions. (A characterisation of all strict identification functions follows invoking Theorem B.1). Patton et al. (2019) and Dimitriadis and Bayer (2019), among many others, utilise these losses for M-estimation of joint semiparametric models.

Due to the recent introduction of ES into the Basel framework as the standard risk measure in banking regulation (Basel Committee, 2016, 2019), there is a fast-growing interest in semiparametric models for ES (jointly with the quantile); see e.g. Patton et al. (2019), Taylor (2019), Bayer and Dimitriadis (2020), Taylor (2020), Dimitriadis et al. (2020), Dimitriadis and Schnaitmann (2020), Meng and Taylor (2020), Barendse (2020), Gerlach and Wang (2020b), and Gerlach and Wang (2020a). Unlike in the two examples concerning multiple quantiles and multiple moments in the previous sections, estimation of semiparametric models for ES can only be performed through a joint model consisting of multiple functionals, which makes it similar to the variance functional. However, unlike the situation of joint mean–variance regression, it turns out that semiparametric M-estimation for $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ can be subject to an efficiency gap.

We assume that all $F \in \mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ are differentiable at the correct α-quantile with strictly positive derivatives there. Also, we assume that the left tail of F is integrable, $\mathbb{E}_F[|Y|\mathbb{1}_{\{Y\le 0\}}] < \infty$, such that ES is finite. Consider the strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function

$$\varphi(y, \xi_1, \xi_2) = \begin{pmatrix} \mathbb{1}_{\{y\le\xi_1\}} - \alpha \\ \xi_2 - \xi_1 + \frac{1}{\alpha}(\xi_1 - y)\mathbb{1}_{\{y\le\xi_1\}} \end{pmatrix}, \tag{4.15}$$

and define the Z-estimator $\widehat{\theta}_{Z,T,\mathbb{A}}$ via (1.6) and (3.1) based on some sequence of instrument matrices $\mathbb{A}$. From Theorem 3.1, we get that the efficient estimator has to fulfil the condition $A^*_{t,C}(X_t, \theta_0) = C D_t(X_t, \theta_0)^\intercal S_t(X_t, \theta_0)^{-1}$ for some deterministic and nonsingular matrix C, where

$$D_t(X_t, \theta_0) = \begin{pmatrix} f_t\big(q_\alpha(X_t, \theta_0)\big)\nabla_\theta q_\alpha(X_t, \theta_0)^\intercal \\ \nabla_\theta e_\alpha(X_t, \theta_0)^\intercal \end{pmatrix} \qquad \text{and} \tag{4.16}$$

$$S_t(X_t, \theta_0) = \begin{pmatrix} \alpha(1-\alpha) & (1-\alpha)\big(q_\alpha(X_t, \theta_0) - e_\alpha(X_t, \theta_0)\big) \\ (1-\alpha)\big(q_\alpha(X_t, \theta_0) - e_\alpha(X_t, \theta_0)\big) & S_{t,22} \end{pmatrix}, \tag{4.17}$$

$$S_{t,22} = \frac{1}{\alpha}\operatorname{Var}_t\left(Y_t \middle| Y_t \le q_\alpha(X_t, \theta_0)\right) + \frac{1-\alpha}{\alpha}\big(e_\alpha(X_t, \theta_0) - q_\alpha(X_t, \theta_0)\big)^2,$$

where straight forward calculations show that the invertibility of $S_t(X_t, \theta_0)$ follows from our assumptions on $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$ which imply that $\operatorname{Var}_t\left(Y_t \middle| Y_t \le q_\alpha(X_t, \theta_0)\right) > 0$. Following Fissler and Ziegel (2016, Theorem 5.2, Corollary 5.5), Fissler and Ziegel (2019a) and our Theorem 2.5, M-estimation of the regression parameters can be carried out (only) by using (strictly) consistent loss functions for the pair $(Q_\alpha, \mathrm{ES}_\alpha)$, which are given by

$$\begin{aligned} \rho_t(y, \xi_1, \xi_2) = {} & \left(\mathbb{1}_{\{y\le\xi_1\}} - \alpha\right) g_t(\xi_1) - \mathbb{1}_{\{y\le\xi_1\}} g_t(y) + \kappa_t(y) \\ & + \phi_t'(\xi_2)\left(\xi_2 - \xi_1 + \frac{1}{\alpha}(\xi_1 - y)\mathbb{1}_{\{y\le\xi_1\}}\right) - \phi_t(\xi_2), \end{aligned} \tag{4.18}$$

where $\xi_1 \mapsto g_t(\xi_1) + \xi_1 \phi_t'(\xi_2)/\alpha$ is (strictly) increasing for each ξ_2, ϕ_t is (strictly) convex and $\mathbb{1}_{\{y \le \xi_1\}} g_t(y)$ and $\kappa_t(y)$ are $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-integrable for all ξ_1.[9] For any sequences $G = (g_t)_{t\in\mathbb{N}}$ and $\Phi = (\phi_t)_{t\in\mathbb{N}}$ of such functions, we denote the M-estimator defined via (1.2) by $\widehat{\theta}_{M,T,G,\Phi}$.

The following theorem establishes that, under certain conditions, the M-estimator of the joint quantile and ES regression model is subject to the *efficiency gap*, i.e., it cannot attain the Z-estimation efficiency bound, and consequently neither the semiparametric efficiency bound.

Theorem 4.3. *Assume that Assumptions (1) – (4) hold, $\widehat{\theta}_{M,T,G,\Phi}$ is asymptotically normal and the following further regularity conditions hold:*

(QES1) *The parameters of the individual models are separated, i.e. the model is of the form $m(X_t, \theta) = \big(q_\alpha(X_t, \theta^q),\ e_\alpha(X_t, \theta^e)\big)^\intercal$, where $\theta = \big(\theta^q, \theta^e\big) \in \Theta \subseteq \mathbb{R}^q$, with $\theta^q \in \mathbb{R}^{q_1}$ and $\theta^e \in \mathbb{R}^{q_2}$ and $q_1 + q_2 = q$.*

(QES2) *For all $t \in \mathbb{N}$, and for all $A \in \mathcal{A}$ with $\mathbb{P}(A) = 1$ there are $q_1 + 1$ mutually different $v_1, \dots, v_{q_1+1} \in \{\nabla_{\theta^q} q_\alpha(X_t(\omega), \theta_0^q) \in \mathbb{R}^{q_1} : \omega \in A\} \subseteq \mathbb{R}^{q_1}$, such that any subset of cardinality q_1 of $\{v_1, \dots, v_{q_1+1}\}$ is linearly independent. The analogue assertion holds for the gradient $\nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$, replacing q_1 by q_2.*

(QES3) *For all $t \in \mathbb{N}$, F_t is differentiable at $q_\alpha(X_t, \theta_0^q)$ with $f_t(q_\alpha(X_t, \theta_0^q)) > 0$ almost surely and $g_t'(\xi_1) + \phi_t'(\xi_2)/\alpha > 0$ and $\phi_t''(\xi_2) > 0$ for all ξ_1, ξ_2.*

Then, the following statements hold:

(A) *Suppose that for all $t \in \mathbb{N}$, $\nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$ almost surely. The M-estimator $\widehat{\theta}_{M,T,G,\Phi}$ attains the Z-estimation efficiency bound in (3.8) if and only if the following five conditions hold:*

$$\exists c_1 > 0\ \forall t \in \mathbb{N}: \quad \operatorname{Var}_t\big(Y_t \big| Y_t \le q_\alpha(X_t, \theta_0^q)\big) = c_1 \big(q_\alpha(X_t, \theta_0^q) - e_\alpha(X_t, \theta_0^e)\big)^2 \quad a.s., \tag{4.19}$$

$$\exists c_2 > 0\ \forall t \in \mathbb{N}: \quad f_t(q_\alpha(X_t, \theta_0^q)) = \frac{c_2}{q_\alpha(X_t, \theta_0^q) - e_\alpha(X_t, \theta_0^e)} \quad a.s., \tag{4.20}$$

$$\exists c_3 > 0\ \forall t \in \mathbb{N}: \quad \phi_t''(e_\alpha(X_t, \theta_0^e)) = \frac{c_3}{\operatorname{Var}_t\big(Y_t \big| Y_t \le q_\alpha(X_t, \theta_0^q)\big)} \quad a.s., \tag{4.21}$$

$$\exists c_4 > 0\ \forall t \in \mathbb{N}\ \exists c_{5,t} \in \mathbb{R}: \quad g_t'(q_\alpha(X_t, \theta_0^q)) = c_4 f_t(q_\alpha(X_t, \theta_0^q)) + c_{5,t} \quad a.s., \tag{4.22}$$

$$\forall t \in \mathbb{N}: \quad \phi_t'(e_\alpha(X_t, \theta_0^e)) = \frac{c_3}{c_1 c_2} f_t(q_\alpha(X_t, \theta_0^q)) - \alpha c_{5,t} \quad a.s. \tag{4.23}$$

(B) *Furthermore, if (4.19), or (4.21), or*

$$\exists c_6 > 0\ \forall t \in \mathbb{N}: \quad g_t'(q_\alpha(X_t, \theta_0^q)) + \phi_t'(e_\alpha(X_t, \theta_0^e))/\alpha = c_6 f_t(q_\alpha(X_t, \theta_0^q)) \quad a.s. \tag{4.24}$$

is violated, then $\widehat{\theta}_{M,T,G,\Phi}$ does not attain the Z-estimation efficiency bound in (3.8).

The general structure of Theorem 4.3 is similar to Theorem 4.2: Statement (A) provides necessary and sufficient conditions as to when the M-estimation and Z-estimation efficiency bounds coincide, using the additional assumption on the model gradients $\nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$.

[9] The positivity assumptions imposed here slightly differ from Dimitriadis and Bayer (2019); Patton et al. (2019). The reason is that the functions g_t and ϕ_t are algebraically not uniquely identified in (4.18). Clearly, replacing $g_t(\xi_1)$ and $\phi_t(\xi_2)$ by $\widetilde{g}_t(\xi_1) = g_t(\xi_1) + c\xi_1$ and $\widetilde{\phi}_t(\xi_2) = \phi(\xi_2) - c\alpha\xi_2$ for some $c \in \mathbb{R}$ leaves the right hand side of (4.18) unchanged, and all subsequent results exhibit this inessential degree of freedom; see (4.22) – (4.24).

Dispensing with the latter condition, (B) provides necessary conditions only. Also, the conditions (QES1) – (QES3) resemble the conditions (DQ1) – (DQ3) and are satisfed for a large class of processes and estimators, see the discussion after Theorem 4.2 and in Patton et al. (2019), Dimitriadis and Bayer (2019).

For the remainder for this section, we assume that the gradient condition $\nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$ holds, putting us in the situation of (A). Then, the core conditions for efficiency of the joint quantile and ES models are given in (4.19) – (4.23), where the conditions (4.19), (4.20) only depend on the underlying process and do not involve g_t and ϕ_t, resembling condition (4.7). These two conditions result from the rather restrictive shape of the class of (strictly) consistent loss functions in (4.18), see Fissler and Ziegel (2016) for details. Section 4.3.2 analyzes the validity of (4.19), (4.20) in location-scale processes with results resembling the ones for double quantile models from the previous section. Given that these conditions hold, efficient M-estimation can be performed by employing suitable choices of g_t and ϕ_t satisfying (4.21) – (4.23), which are further discussed in Section 4.3.1 and which resemble conditions (4.8) and (4.9).

Conditions (4.19) – (4.23) and (4.24) nicely illustrate the concordance with mean and quantile regression models. Condition (4.24) (which can be split into (4.22) and (4.23) under the equality of the model gradients) is closely related to the efficient choice for semiparametric quantile models, see Komunjer and Vuong (2010b,a), and Section 4.2 of this article. However, in contrast to classical quantile regression, it is important to notice that given (4.19) and (4.20) hold, the choice $g_t(z) = 0$ (resulting from $c_4 = 0$ and $c_{5,t} = 0$) facilitates efficient estimation through a suitable choice of the function ϕ_t. Moreover, condition (4.21) resembles the classical condition of efficient least squares estimation of Gourieroux et al. (1984), where the second derivative of the Bregman function is proportional to the reciprocal of the conditional variance. As ES is a *tail* expectation, one also needs to consider the *tail variance* in (4.21).

Barendse (2020) considers a two-step estimation procedure, and a related *two-step efficiency bound*, for semiparametric quantile and ES models, which we further discuss and relate to the results of Theorem 4.3 in Appendix D.

4.3.1. Efficient Estimation of Joint Semiparametric Quantile and ES Models

We discuss feasible choices for g_t and ϕ_t satisfying (4.21) – (4.23) and (QES3) in order to facilitate efficient M-estimation for semiparametric joint quantile and ES models based on the results of Theorem 4.3. To this end, we assume that (4.19) and (4.20) hold for the underlying process and defer a discussion of these conditions to Section 4.3.2. An obvious solution satisfying (4.21) – (4.23) is given by

$$\begin{aligned} g_t^{\text{eff1}}(\xi_1) &= d_1 F_t(\xi_1), && \text{for all} \quad \xi_1 > e_\alpha(X_t, \theta_0^e), \\ \phi_t^{\text{eff1}}(\xi_2) &= -d_2 \log\big(q_\alpha(X_t, \theta_0^q) - \xi_2\big) && \text{for all} \quad \xi_2 < q_\alpha(X_t, \theta_0^q), \end{aligned} \tag{4.25}$$

almost surely for all $t \in \mathbb{N}$ and for some constants $d_1 \geq 0$ and $d_2 > 0$, which we refer to as the *first pseudo-efficient* choices. Motivated through the condition

$$\phi_t''(e_\alpha(X_t, \theta_0^e)) = c\Big(\operatorname{Var}_t\big(Y_t \big| Y_t \leq q_\alpha(X_t, \theta_0^q)\big) + (1-\alpha)\big(e_\alpha(X_t, \theta_0^e) - q_\alpha(X_t, \theta_0^q)\big)^2 \Big)^{-1} \tag{4.26}$$

for some constant $c > 0$, given in (F.11) in the proof of Theorem 4.3 and in the two-step estimation efficiency bound of Barendse (2020), a *second pseudo-efficient* choice, satisfying (4.21) –

(4.23), is given by

$$\begin{aligned} g_t^{\text{eff2}}(\xi_1) &= 0, \qquad \text{and} \\ \phi_t^{\text{eff2}}(\xi_2) &= \frac{d_3(q_t - \xi_2)}{\sqrt{(1-\alpha)v_t}} \arctan\left(\frac{\sqrt{1-\alpha}(q_t-\xi_2)}{\sqrt{v_t}}\right) + \xi_2 \frac{\pi d_3(1+d_4)}{2\sqrt{(1-\alpha)v_t}} \\ &\quad - \frac{d_3}{2(1-\alpha)} \log\left(v_t + (1-\alpha)(q_t-\xi_2)^2\right), \quad \text{for all} \quad \xi_2 < q_t, \end{aligned} \tag{4.27}$$

almost surely for some constants $d_3 > 0$, $d_4 \geq 0$, where $v_t = \text{Var}_t\left(Y_t \middle| Y_t \leq q_\alpha(X_t, \theta_0^q)\right)$ and $q_t = q_\alpha(X_t, \theta_0^q)$. For this choice, it holds that

$$\phi_t'^{\text{eff2}}(\xi_2) = -\frac{d_3}{\sqrt{(1-\alpha)v_t}} \arctan\left(\frac{\sqrt{1-\alpha}(q_t-\xi_2)}{\sqrt{v_t}}\right) + \frac{\pi d_3(1+d_4)}{2\sqrt{(1-\alpha)v_t}} > 0,$$

and $\phi_t''^{\text{eff2}}(\xi_2) = d_3\big(v_t + (1-\alpha)(q_t-\xi_2)^2\big)^{-1} > 0$, for all $\xi_2 < q_t$.

This illustrates that, given that (4.19) and (4.20) hold, there exist multiple essentially different and efficient M-estimators. Furthermore, if (4.19) or (4.20) do not hold jointly, Theorem 4.3 cannot be employed for a statement on efficiency of different M-estimators and it is generally unclear which choices of g_t and ϕ_t result in the most efficient estimator. We analyze this numerically in situations of a location-scale process with heteroskedastic innovation distributions in the simulation study in Section 5. The results there lead to the conjecture that there is also an efficiency gap in models with joint parameters.

As it is common for efficient semiparametric estimation (cf. Gourieroux et al., 1984, Komunjer and Vuong, 2010b,a), the efficient choice depends on the knowledge of the true parameter vector θ_0 and further unknown quantities such as the distribution function $F_t\big(q_\alpha(X_t, \theta_0^q)\big)$ and the quantile-truncated variance $\text{Var}_t\left(Y_t \middle| Y_t \leq q_\alpha(X_t, \theta_0^q)\right)$. In practice, one usually applies a two-step estimation approach where the unknown quantities in the efficient choices are substituted by consistent estimates. Notably, the pseudo-efficient M-estimators based on the first choices $g_t(\xi_1) = 0$ and $\phi_t(\xi_2)$ given in (4.25) are remarkably feasible in the sense that they only require a first-step estimate of the quantile-specific parameters. This is considerably easier than the required nonparametric first-step estimators of the conditional variance or the conditional distribution function in efficient M-estimation of mean and quantile regressions.

A further interesting fact arises from a comparison of (4.25) to the predominantly used loss functions with positively homogeneous loss differences of degree zero (Nolde and Ziegel, 2017), which are essentially characterised by the choices

$$g_t(\xi_1) = 0 \qquad \text{and} \qquad \phi_t(\xi_2) = -\log(-\xi_2), \qquad \xi_2 < 0. \tag{4.28}$$

Patton et al. (2019) build their M-estimation approach on these choices and Dimitriadis and Bayer (2019) numerically show that such M-estimators appear to be relatively efficient.

Comparing the choice $g_t(\xi_1) = 0$ to the efficient choice in (4.25) illustrates the elegance of the parsimonious choice $d_1 = 0$. By further comparing the choices of ϕ_t in (4.25) and (4.28), we see that the zero-homogeneous loss function only deviates from the pseudo-efficient choice in (4.25) through the translation by $q_\alpha(X_t, \theta_0^q)$. This justifies the choice of Patton et al. (2019) ex post and theoretically explains the good numerical performance observed by Bayer and Dimitriadis (2020). While the zero-homogeneous choice requires strictly negative values for the conditional ES, employing the closely related efficient choice in (4.25) makes this condition redundant and instead, we only have to impose the natural condition that the conditional ES be smaller than the conditional quantile. The choices coincide if $q_\alpha(X_t, \theta_0) = 0$. Interestingly, when $d_1 = 0$, (4.25) also constitutes a strictly consistent loss with zero-homogeneous loss differences, when allowing the (itself 1-homogenous) quantile as an input parameter. This does not contradict Nolde and Ziegel (2017) as they naturally do not allow the true quantile as an input parameter.

4.3.2. Model Examples: Joint Quantile and ES Regression

In this section, we discuss the attainability of the process conditions (4.19) and (4.20), which are necessary for the M-estimator to match the Z-estimation efficiency bound. For this, we consider the same generalised location-scale process $Y_t = \mu(X_t, \eta_0) + \sigma(X_t, \eta_0)u_t$ as in (4.11), where $\mu(X_t, \eta_0)$ and $\sigma(X_t, \eta_0)$ are parametric models for the conditional mean and scale. The innovations $(u_t)_{t\in\mathbb{N}}$ are themselves independent, independent of $(X_t)_{t\in\mathbb{N}}$, and satisfy the semi-parametric stationarity condition that $z_\alpha = F_{u_t}^{-1}(\alpha)$ and $\zeta_\alpha = \mathrm{ES}_\alpha(u_t)$ are time-independent, such that Assumption (2) holds. Apart from that, the innovations may be heterogeneously distributed.

For this process, we get that $Q_\alpha(Y_t|X_t) = \mu(X_t, \eta_0) + \sigma(X_t, \eta_0)z_\alpha$ and $\mathrm{ES}_\alpha(Y_t|X_t) = \mu(X_t, \eta_0) + \sigma(X_t, \eta_0)\zeta_\alpha$, and furthermore $\mathrm{Var}_t\big(Y_t \big| Y_t \le q_\alpha(X_t, \theta_0^q)\big) = \mathrm{Var}_t\big(u_t \big| u_t \le z_\alpha\big)\sigma(X_t, \eta_0)^2$, and $f_t\big(q_\alpha(X_t, \theta_0^q)\big) = f_{u_t}(z_\alpha)/\sigma(X_t, \eta_0) = f_{u_t}(z_\alpha)(z_\alpha - \zeta_\alpha)/(q_\alpha(X_t, \theta_0^q) - e_\alpha(X_t, \theta_0^e))$. Thus, for stationary innovations $(u_t)_{t\in\mathbb{N}}$, the quantities $\mathrm{Var}_t\big(u_t \big| u_t \le z_\alpha\big)$ and $f_{u_t}(z_\alpha)$ are constant, which implies that the conditions (4.19) and (4.20) are satisfied, and hence, any M-estimator based on choices for g_t and ϕ_t satisfying (4.21)–(4.23) attains the Z-estimation efficiency bound if $\nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$ holds.

Similarly to Section 4.2, we are able to construct processes which generate an *efficiency gap* by considering time-varying innovation distributions. E.g., we consider independently and Student's t-distributed innovations $u_t \sim t_{\nu_t}(\mu_t, \sigma_t)$ with time-varying degrees of freedom ν_t and

$$\mu_t = Q_\alpha(t_{\nu_1}) - \sigma_t Q_\alpha(t_{\nu_t}) \qquad \text{and} \qquad \sigma_t = \frac{Q_\alpha(t_{\nu_1}) - \mathrm{ES}_\alpha(t_{\nu_1})}{Q_\alpha(t_{\nu_t}) - \mathrm{ES}_\alpha(t_{\nu_t})}, \tag{4.29}$$

in order to satisfy the quantile-ES stationarity condition. For this process, it still holds that $\mathrm{Var}_t\big(Y_t|Y_t \le q_\alpha(X_t, \theta_0)\big) = \sigma^2(X_t, \eta_0)\,\mathrm{Var}\big(u_t|u_t \le z_\alpha\big)$, as u_t is independent of X_t. However, the quantity $\mathrm{Var}\big(u_t|u_t \le z_\alpha\big)$ is generally time-varying, and consequently, this violates (4.19).

5. Numerical Illustration of the Efficiency Gap

In this section, we numerically illustrate the efficiency gap for joint semiparametric models for the quantile and ES. For results on double quantile models, we refer to Appendix E. Following (3.2)–(3.4), we use a Monte Carlo approach to approximate the expectations (over the covariates) in the closed-form formulas for the asymptotic covariances. We use 1000 simulation replications each consisting of a sample size of $T = 2000$.

For joint quantile and ES models with separated model parameters, we use the process in (4.11) and utilize parametric choices which result in strictly negative ES values,

$$X_t \overset{iid}{\sim} 3\,\mathrm{Beta}(3, 1.5), \quad \mu(X_t, \eta_0) = -1 - 0.5X_t, \quad \text{and} \quad \sigma(X_t, \eta_0) = 0.5 + 0.5X_t. \tag{5.1}$$

For the model innovations u_t, we choose the following two specifications:

(a) $u_t \overset{iid}{\sim} \mathcal{N}(0, 1)$,

(b) $u_t \sim t_{\nu_t}(\mu_t, \sigma_t)$ with time-varying degrees of freedom, $\nu_t = 3\,\mathbb{1}_{\{t \le T/2\}} + 100\,\mathbb{1}_{\{t > T/2\}}$, where μ_t and σ_t are given in (4.29).

Mathematically, these choices are motivated through the theoretical considerations of Section 4.3 that for location-scale models with i.i.d. residuals, the M-estimator is able to attain the Z-estimation efficiency bound, while conversely, it cannot do so for heterogeneously distributed innovations. Empirically, the particular scenario of (b) can be motivated by a breakpoint model for the degree of heavy tailedness of the innovations: A period of stress (first part of the sample) exhibiting heavy tails, is followed by a relatively calm period (second part of the sample), which is resembled an innovation-distribution with considerably lighter tails. We estimate the following linear models with separated parameters,

$$q_\alpha(X_t, \theta) = \theta^{(1)} + \theta^{(2)} X_t, \quad \text{and} \quad e_\alpha(X_t, \theta) = \theta^{(3)} + \theta^{(4)} X_t, \tag{5.2}$$

which satisfy the conditions of Theorem 4.3.

We further consider linear models with *joint* model parameters where the conditions of Theorem 4.3 do not hold in order to assess efficient estimation of quantile–ES models beyond the model classes considered in Theorem 4.3. For this, we use a slightly modified parameterisation of the process by using $\sigma(X_t, \eta_0) = 0.5X_t$, which implies that $Q_\alpha(Y_t|X_t) = -1 + (0.5z_\alpha - 0.5)X_t$ and $\mathrm{ES}_\alpha(Y_t|X_t) = -1 + (0.5\zeta_\alpha - 0.5)X_t$. Hence, we use the (correctly specified) *joint intercept models*

$$q_\alpha(X_t, \theta) = \theta^{(1)} + \theta^{(2)} X_t, \quad \text{and} \quad e_\alpha(X_t, \theta) = \theta^{(1)} + \theta^{(3)} X_t. \tag{5.3}$$

We consider the quantile and ES at joint probability levels $\alpha \in \{1\%, 2.5\%, 10\%\}$ and for a detailed list of the employed functions g_t and ϕ_t, see the first two columns of Table 1. For g_t, we use the two (pseudo-efficient) choices $g_t(\xi_1) = 0$ and $g_t(\xi_1) = F_t(\xi_1)$ coupled with the following choices of ϕ_t. The first two choices of ϕ_t correspond to sub-optimal choices as already noticed by Dimitriadis and Bayer (2019), whereas the next choice coincides with the ubiquitous zero-homogeneous choice $\phi_t(\xi_2) = -\log(-\xi_2)$. The latter two choices ϕ_t^{eff1} and ϕ_t^{eff2} are the pseudo-efficient choices given in (4.25) and (4.27).

Panel A of Table 1 presents the approximated parameter standard deviations for $\alpha = 2.5\%$ and for the separated parameter models in (5.2). The results confirm the theoretical considerations of Section 4.3: the M-estimator based on either of the pseudo-efficient choices, ϕ_t^{eff1} and ϕ_t^{eff2}, attains the Z-estimation efficiency bound for location-scale models with homoskedastic innovations, while there is an efficiency gap for heteroskedastic innovation distributions with a not inessential magnitude of up to 15%. Table 5 (in Appendix E) reports additional results for $\alpha = 1\%$ and $\alpha = 10\%$, which show that the efficiency gap seems to be more pronounced for small(er) probability levels, corresponding to the most important cases for the risk measures VaR and ES.

As indicated by (4.25), our simulation results confirm that, though unintuitive at first sight, efficient M-estimation in the homoskedastic case can be accomplished by both, the traditional efficient choice of quantile regression, $g_t(\xi_1) = F_t(\xi_1)$, and by the zero-function $g_t(\xi_1) = 0$. Furthermore, both pseudo-efficient choices ϕ^{eff1} and ϕ^{eff2} are able to attain the efficiency bound in the homoskedastic setting for separated parameter models. However, their performance differs for heteroskedastic models, where in this instance, the choice ϕ^{eff2} delivers more efficient ES estimates but at the same time slightly less efficient quantile estimates. The function $\phi_t(\xi_2) = -\log(-\xi_2)$ performs almost as good as the pseudo-efficient choices throughout all considered designs, which is not surprising given its similar form to ϕ^{eff1}.

Panel B of Table 1 presents results for the models with joint parameters given in (5.3). While

Table 1.: Asymptotic Standard Deviations of Joint Quantile and ES Models

$g_t(\xi_1)$	$\phi_t(\xi_2)$	(a) Homoskedastic θ_1	θ_2	θ_3	θ_4	(b) Heteroskedastic θ_1	θ_2	θ_3	θ_4
		Panel A: Models with Separated Parameters							
0	$\exp(\xi_2)$	12.249	7.588	13.101	7.628	84.231	64.008	112.745	84.790
$F_t(\xi_1)$	$\exp(\xi_2)$	11.772	7.028	13.101	7.628	55.675	28.526	112.745	84.790
0	$F_{\text{Log}}(\xi_2)$	12.128	7.485	12.950	7.507	83.881	63.663	112.118	84.135
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	11.663	6.941	12.950	7.507	55.442	28.392	112.118	84.135
0	$-\log(-\xi_2)$	9.949	5.480	11.940	6.576	36.254	17.773	71.900	39.615
$F_t(\xi_1)$	$-\log(-\xi_2)$	9.949	5.480	11.940	6.576	36.237	17.765	71.900	39.615
0	$\phi_t^{\text{eff1}}(\xi_2)$	9.942	5.476	11.907	6.559	36.242	17.766	71.832	39.579
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	9.942	5.476	11.907	6.559	36.242	17.766	71.832	39.579
0	$\phi_t^{\text{eff2}}(\xi_2)$	9.942	5.476	11.907	6.559	37.837	18.352	67.775	37.342
Barendse Bound		10.369	5.701	11.907	6.559	35.397	17.337	67.775	37.342
Efficiency Bound		9.942	5.476	11.907	6.559	31.248	15.290	59.332	33.933
		Panel B: Models with Joint Parameters							
0	$\exp(\xi_2)$	3.929	2.855	2.919		26.757	23.011	23.775	
$F_t(\xi_1)$	$\exp(\xi_2)$	3.897	2.800	2.898		25.431	19.266	22.531	
0	$F_{\text{Log}}(\xi_2)$	3.935	2.825	2.902		26.469	22.287	23.202	
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	3.900	2.772	2.880		25.111	18.625	21.954	
0	$-\log(-\xi_2)$	4.035	2.706	2.847		21.331	11.515	15.145	
$F_t(\xi_1)$	$-\log(-\xi_2)$	4.027	2.702	2.843		21.288	11.499	15.129	
0	$\phi_t^{\text{eff1}}(\xi_2)$	3.832	2.607	2.753		21.862	12.212	15.717	
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	3.827	2.604	2.751		21.676	12.129	15.633	
0	$\phi_t^{\text{eff2}}(\xi_2)$	3.758	2.570	2.718		22.311	12.117	15.401	
Efficiency Bound		3.676	2.529	2.680		12.586	8.144	11.456	

This table presents the (approximated) asymptotic standard deviations for semiparametric joint quantile and ES models at joint probability level of 2.5% for various choices of M-estimators together with the Z-estimation efficiency bound. Panel A reports results for the models with separated parameters given in (5.2) while Panel B considers the joint intercept models given in (5.3). The two considered residual distributions are presented in the two vertical panels of the table. The line "Barendse Bound" in Panel A refers to the two-step efficiency bound of Barendse (2020) discussed in Appendix D and is reported here for completeness.

the Z-estimation efficiency bound is still valid, it cannot be attained by any of the M-estimators utilised in the simulation study, even in the homoskedastic case, for any of the chosen pseudo-efficient choices. This implies that the joint quantile and ES models exhibit an even more general efficiency gap which goes beyond the model class considered in Theorem 4.3. This holds similarly for the heteroskedastic case, where the efficiency gap becomes quantitatively much larger: the standard deviations of the pseudo-efficient choices are up to almost the double of the standard deviations at the efficiency bound.

As in the heteroskedastic case of Panel A, the second pseudo-efficient choice slightly outperforms the first one also for this example of joint parameter models. Finally, among the considered M-estimators, the ubiquitous zero-homogeneous choice $g_t(\xi_1) = 0$ and $\phi_t(\xi_2) = -\log(-\xi_2)$ performs relatively well and even outperforms both pseudo-efficient choices for the heteroskedastic innovation distributions and the joint parameter models of Panel B. This is especially remarkable given that, in contrast to the pseudo-efficient choices, it does not require any pre-estimates.

6. Conclusion

In this article, we show the existence of an "efficiency gap" between the M- and Z-estimator for semiparametric modelling of multivariate functionals. Such a gap does not exist in the setup for univariate functionals, where the *efficient* M-estimator can attain the Z-estimation (semiparametric) efficiency bound. We show that for certain processes and pairs of functionals, the M-estimator cannot attain the semiparametric efficiency bound (even in its practically infeasible form).

For this, we derive characterization results for the classes of (consistent) M-estimators by formalizing the relationship between (consistent) M-estimation and strictly consistent loss functions in Theorem 2.5, and by drawing on existing characterization results of (strictly) consistent loss functions from the decision theoretic literature (Gneiting, 2011a; Fissler and Ziegel, 2016). Furthermore, we show that the "optimal instrument matrix" of Hansen (1985), Chamberlain (1987), Newey (1993) in efficient GMM- / Z-estimation based on conditional moment restrictions is not only a sufficient, but also a necessary condition for efficient estimation.

We illustrate this gap for two pairs of semiparametric models. First, we consider semiparametric quantile models for two quantile levels simultaneously, which are motivated through VaR forecasting at multiple (VaR) levels, and through prediction intervals, e.g. for COVID-19 modelling and forecasting. For these, we identify conditions where no efficient choice of the M-estimator exists – even adaptions of the efficient semiparametric M-estimator (for one quantile level) of Komunjer and Vuong (2010b,a) do not attain the bound (under these conditions).

Second, we examine efficient M-estimation for joint quantile and ES models of Patton et al. (2019), which recently attracts a lot of attention due to the introduction of the ES into the Basel regulatory framework (Basel Committee, 2016, 2019) and the joint elicitability of the VaR and ES (Fissler and Ziegel, 2016). As for the double quantile case, joint VaR and ES models are also subject to the efficiency gap in the general case. Through restricting attention to location-scale models, which is the most popular model class in the financial literature, we derive two pseudo efficient choices – they attain the efficiency bound in (and only in) location-scale models. We find that the first pseudo-efficient choice is "surprisingly feasible", and contrary to efficient estimators for semiparametric models for many other functionals, does require little pre-estimates. Furthermore, it is closely related to the zero-homogeneous choice, which can explain the good empirical (efficient) performance of the latter. Open points are optimal choices of the underlying objective function for the M-estimator for general processes, and consequently, semiparametric M-estimation efficiency bounds.

While we illustrate the theory for general multivariate functionals on two examples – two quantiles and the ES jointly with a quantile – our setup allows evaluation of efficient semiparametric M-estimation for further multivariate functionals such as multiple expectiles or the inter-quantile expectation (or Range Value at Risk), together with its associated quantiles. Furthermore, it allows for the consideration of more than two functionals (e.g. multiple quantiles), where our results directly carry over.

Acknowledgements

We are very grateful to Jana Hlavinová for a careful proofreading and valuable feedback on an earlier version of this paper, which helped to improve the paper. We further thank Sander Barendse for a fruitful discussion of the results of the paper.

Appendix

A. Further implications of the gap: Equivariance properties

Patton (2011) and Nolde and Ziegel (2017) provide arguments for the usage of homogeneous loss functions for forecast comparison and ranking. More generally, Fissler and Ziegel (2019b) advocate for loss functions that respect equivariance properties of the functional of interest. Besides homogeneity, a major equivariance property of interest is translation equivariance, or – more generally speaking – linear equivariance; see Fissler and Ziegel (2019b). Again, we focus on two interesting pairs of functionals, (mean, variance) and $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$. For any random variable Y with finite second moment and any scalar $c \in \mathbb{R}$, the following identities hold

$$\begin{aligned} \big((\mathrm{VaR}_\alpha(Y+c), \mathrm{ES}_\alpha(Y+c)\big) &= \big((\mathrm{VaR}_\alpha(Y)+c, \mathrm{ES}_\alpha(Y)+c\big), \\ \big(\mathbb{E}[Y+c], \mathrm{Var}(Y+c)\big) &= \big(\mathbb{E}[Y]+c, \mathrm{Var}(Y)\big)\,. \end{aligned} \tag{A.1}$$

Suppose one is to model the functional $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ with a parametric model (possibly with joint model parameters) of the form $m(X,\theta) = \big(q_\alpha(X,\theta), e_\alpha(X,\theta)\big)$, where $\theta = \big(\theta^{(1)}, \dots, \theta^{(q)}\big) \in \Theta \subseteq \mathbb{R}^q$, with intercept parameters, say

$$\begin{pmatrix} q_\alpha(X,\theta) \\ e_\alpha(X,\theta) \end{pmatrix} = \begin{pmatrix} \theta^{(1)} + \widetilde{q}_\alpha(X, \theta^{(3)}, \dots, \theta^{(q)}) \\ \theta^{(2)} + \widetilde{e}_\alpha(X, \theta^{(3)}, \dots, \theta^{(q)}) \end{pmatrix}.$$

Then, under Assumption (1), the correctly specified parameter θ_0 has the following equivariance property for $(Y,X) \in \mathcal{Z}$ and $c \in \mathbb{R}$ such that $(Y+c, X) \in \mathcal{Z}$:

$$\theta_0^{(j)}(F_{(Y+c,X)}) = \begin{cases} \theta_0^{(j)}(F_{(Y,X)}) + c, & \text{for } j = 1,2, \\ \theta_0^{(j)}(F_{(Y,X)}), & \text{for } j = 3,\dots,q. \end{cases} \tag{A.2}$$

Similar results apply to the pair (mean, variance), where, of course, the intercept transformation only appears in the mean-component.

Similarly, given data $(\boldsymbol{Y}, \boldsymbol{X}) = (Y_t, X_t)_{t=1,\dots,T}$, it would be desirable to find a similar translation equivariance property for an estimator $\widehat{\theta}_T = \widehat{\theta}_T(\boldsymbol{Y}, \boldsymbol{X})$:

$$\widehat{\theta}_T^{(j)}(\boldsymbol{Y}+c, \boldsymbol{X}) = \begin{cases} \widehat{\theta}_T^{(j)}(\boldsymbol{Y}, \boldsymbol{X}) + c, & \text{for } j = 1,2, \\ \widehat{\theta}_T^{(j)}(\boldsymbol{Y}, \boldsymbol{X}), & \text{for } j = 3,\dots,q. \end{cases} \tag{A.3}$$

Under Assumption (1) of a correctly specified model, (A.3) holds for the probability limit of any consistent estimator. However, in finite samples or under model misspecification, it may well fail unless there is some additional structure in the estimator. For example, the OLS-estimator clearly satisfies (A.2) and (A.3), relying on the fact that the squared loss $\rho(y,\xi) = \frac{1}{2}(y-\xi)^2$ is translation invariant. Also, the corresponding Z-estimator is translation equivariant, since the standard identification function $\varphi(y,\xi) = y - \xi$ is translation invariant and the instrument matrix $A(X,\theta) = X$ is independent of θ.

It turns out that both two-dimensional functionals in (A.1) possess strict identification functions that respect the respective equivariance properties described there, namely

$$\varphi_{(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)}(y, \xi_1, \xi_2) = \begin{pmatrix} \mathbb{1}\{y \le \xi_1\} - \alpha \\ \xi_2 + \frac{1}{\alpha}\mathbb{1}\{y \le \xi_1\}(\xi_1 - y) - \xi_1 \end{pmatrix}, \quad \varphi_{(\mathbb{E}, \mathrm{Var})}(y, \xi_1, \xi_2) = \begin{pmatrix} \xi_1 - y \\ \xi_2 - (y-\xi_1)^2 \end{pmatrix}.$$

Using instrument matrices which are independent of θ, they induce Z-estimators which obey the translation equivariance in their intercept components. However, Propositions 4.9 and 4.10 in Fissler and Ziegel (2019b) ascertain that for both functional pairs, there are no strictly consistent loss function with these equivariance properties – at least under general and realistic assumptions. This rules out the existence of corresponding M-estimators with this property – another manifestation of the gap between these two classes of estimators.

B. Characterisation of identification functions

For the exactly identified case where the dimension of the identification function ℓ coincides with the dimension of the action domain k, and under some richness conditions on the class of distributions $\mathcal{F}$ given in Assumptions (5) and (6), one can derive a characterisation of the class of strict $\mathcal{F}$-identification functions for a given functional $\Gamma\colon \mathcal{F} \to \Xi$. Even though the proof borrows arguments from Osband's principle (Osband, 1985; Gneiting, 2011a; Fissler and Ziegel, 2016) the assertion has recently been stated in the PhD thesis Fissler (2017, Proposition 3.2.1).

Assumption (5). Let $\mathcal{F}$ be a convex class of distributions on $\mathbb{R}^d$ such that for every $\xi \in \operatorname{int}(\Xi) \subseteq \mathbb{R}^k$ there are $F_1, \dots, F_{k+1} \in \mathcal{F}$ satisfying $0 \in \operatorname{int}\left(\operatorname{conv}\left(\{\bar\varphi(F_1,\xi), \dots, \bar\varphi(F_{k+1},\xi)\}\right)\right)$.

Assumption (6). For every $y \in \mathbb{R}^d$ there exists a sequence $(F_n)_{n\in\mathbb{N}}$ of distributions $F_n \in \mathcal{F}$ that converges weakly to the Dirac-measure δ_y and a compact set $K \subset \mathbb{R}^d$ such that the support of F_n is contained in K for all n.

Assumption (7). Suppose that φ and φ' are locally bounded jointly in their two arguments. Moreover, suppose that the complement of the set

$$C := \{(y,\xi) \in \mathbb{R}^d \times \operatorname{int}(\Xi) \,|\, \varphi(\cdot,\xi) \text{ and } \varphi'(\cdot,\xi) \text{ are continuous at the point } y\}$$

has $(d+k)$-dimensional Lebesgue measure zero.

Assumption (5) corresponds to Assumption (V1), Assumption (6) to Assumption (F1), and Assumption (7) to Assumption (VS1) in Fissler and Ziegel (2016), respectively. We only add the local boundedness in Assumption 7; see the discussion in Brehmer (2017) and the erratum Fissler and Ziegel (2019a).

Theorem B.1 (Proposition 3.2.1 in Fissler (2017))**.** *Let $\Gamma\colon \mathcal{F} \to \Xi \subseteq \mathbb{R}^k$ be a functional with a strict $\mathcal{F}$-identification function $\varphi\colon \mathbb{R}^d \times \Xi \to \mathbb{R}^k$. Then the following two assertions hold:*

(i) *If $h\colon \Xi \to \mathbb{R}^{k\times k}$ is a matrix-valued function with $\det(h(\xi)) \neq 0$ for all $\xi \in \Xi$, then $(y,\xi) \mapsto h(\xi)\varphi(y,\xi)$ is also a strict $\mathcal{F}$-identification function for Γ.*

(ii) *Let φ satisfy Assumption (5) and let $\varphi'\colon \mathbb{R}^d \times \Xi \to \mathbb{R}^k$ be an $\mathcal{F}$-identification function for Γ. Then there is a matrix-valued function $h\colon \operatorname{int}(\Xi) \to \mathbb{R}^{k\times k}$ such that*

$$\bar\varphi'(F,\xi) = h(\xi)\bar\varphi(F,\xi)$$

for all $\xi \in \operatorname{int}(\Xi)$ and for all $F \in \mathcal{F}$. If φ' is a strict $\mathcal{F}$-identification function for Γ and it also satisfies Assumption (5), then additionally $\det(h(\xi)) \neq 0$ for all $\xi \in \operatorname{int}(\Xi)$. If the integrated identification functions $\bar\varphi(F,\cdot)$ and $\bar\varphi'(F,\cdot)$ are continuous, then also h is continuous, which implies that either $\det(h(\xi)) > 0$ for all $\xi \in \operatorname{int}(\Xi)$ or $\det(h(\xi)) < 0$ for all $\xi \in \operatorname{int}(\Xi)$. Moreover, if $\mathcal{F}$ satisfies Assumption (6) and φ, φ' satisfy Assumption (7) it even holds that

$$\varphi'(y,\xi) = h(\xi)\varphi(y,\xi) \tag{B.1}$$

for almost all $(y,\xi) \in \mathbb{R}^d \times \mathrm{int}(\Xi)$.

Remark B.2. It is possible to use a modification of part (i) of Theorem B.1 to construct identification functions for the unidentified case ($\ell < k$) and the overidentified case ($\ell > k$). One can simply work with a function h taking values in $\mathbb{R}^{\ell\times k}$. However, for $\ell < k$ this approach generally leads to an identification function that fails to be strict. In contrast, for $\ell > k$ the resulting identification function will be strict provided that h has full rank. For example, this leads to the strict 2-dimensional identification function for the mean functional $\psi(y,\xi) = \big(\xi - y, (\xi^2+1)(\xi-y)\big)^\intercal$. On the other hand, part (ii) in Theorem B.1 fails to be applicable for a characterisation of overidentified strict identification functions constructed this way. The very reason is that Assumption (5) will not hold (upon replacing k with ℓ).

The following example illustrates possible choices of instrument matrices in a linear mean regression context.

Example B.3. Let $k=1$, $q=p\geq 1$ and $\mathcal{X}$ be such that $\mathbb{E}[XX^\intercal]$ has full rank for all $X\in\mathcal{X}$. Then define

$$\mathcal{F}_{\mathcal{Z}} = \{F_{m(X,\theta_0)+\varepsilon,X} \mid X\in\mathcal{X},\ \theta_0\in\Theta=\mathbb{R}^q,\ \mathbb{E}[\varepsilon|X]=0\}, \tag{B.2}$$

where $m(X,\theta) = X^\intercal\theta$. Clearly, $\Gamma(F_{Y|X}) := \int y \,\mathrm{d}F_{Y|X}(y) = X^\intercal\theta_0$. The condition that $\mathbb{E}[XX^\intercal]$ has full rank implies that the model m uniquely identifies the parameter θ_0 such that Assumption (1) holds. Indeed, for any $\theta'\neq\theta$ it holds that $0 < (\theta-\theta')^\intercal\mathbb{E}\big[XX^\intercal\big](\theta-\theta') = \mathbb{E}\big[\|m(X,\theta) - m(X,\theta')\|^2\big]$. Employing the canonical identification function for the mean, we obtain a strict conditional $\mathcal{F}_{\mathcal{Z}}$-identification function $\varphi\big(Y,m(X,\theta)\big) = m(X,\theta) - Y = X^\intercal\theta - Y$. Indeed $\mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)|X\big] = X^\intercal(\theta-\theta_0)$. If $\theta\neq\theta_0$ and $X^\intercal(\theta-\theta_0)$ were 0 a.s., then $\mathbb{E}[XX^\intercal](\theta-\theta_0)=0$, violating the full rank property of $\mathbb{E}[XX^\intercal]$.[10]
Now, choosing an instrument matrix $A(X,\theta)$ such that $\mathbb{E}[A(X,\theta)X^\intercal]$ has full rank for all θ is a sufficient and necessary condition to ensure that $\psi_A(Y,X,\theta) = A(X,\theta)\varphi(Y,m(X,\theta))$ is a strict *unconditional* $\mathcal{F}_{\mathcal{Z}}$-identification function for θ_0. Indeed, we obtain $\mathbb{E}[\psi_A(Y,X,\theta)] = \mathbb{E}\big[A(X,\theta)X^\intercal\big](\theta-\theta_0)$. In particular, the choice $A(X,\theta)=X$ yields a strict unconditional $\mathcal{F}_{\mathcal{Z}}$-identification function for θ_0. □

The following proposition generalises the rationale described in Example B.3 from the well-known situation of linear models, establishing a sufficient condition on the instrument matrix $A(X,\theta)$ such that $A(X,\theta)\varphi(Y,m(X,\theta))$ becomes a strict identification function for θ_0.

Proposition B.4. *Under Assumption (1), let* $\varphi\colon\mathbb{R}\times\mathbb{R}^k\to\mathbb{R}^k$ *be a strict* $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$*-identification function. Let* $A\colon\mathbb{R}^p\times\Theta\to\mathbb{R}^{q\times k}$ *be an instrument matrix such that* $\mathbb{E}\big[A(X,\theta)D(X,\theta')\big]$ *has full rank, where*

$$D(X,\theta') = \nabla_\theta\mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)|X\big]\,\big|_{\theta=\theta'}$$

for all $(Y,X)\in\mathcal{Z}$ *and for all* $\theta,\theta'\in\Theta$ *such that there is a* $\lambda\in[0,1]$ *with* $\theta' = (1-\lambda)\theta_0+\lambda\theta$*. Then* $A(x,\theta)\varphi\big(y,m(x,\theta)\big)$ *is a strict unconditional* $\mathcal{F}_{\mathcal{Z}}$*-identification function for* θ_0*.*

The following proposition takes the angle of 'reverse engineering' establishing a counterpart to Theorem 2.5 (iv): When an unconditional strict identification function is of the form (2.7) it establishes a sufficient condition on the instrument matrix A to ensure that φ is a conditional strict identification function.

[10] However, $\varphi\big(Y,m(X,\theta)\big)$ is in general not a *strict unconditional* $\mathcal{F}_{\mathcal{Z}}$-identification function for θ_0. It could be the case, for example, that there is an $X\in\mathcal{X}$ with $\mathbb{E}[X]=0$. Then we one obtain $\mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)\big] = \mathbb{E}[X^\intercal(\theta-\theta_0)] = 0$ for all θ.

Proposition B.5. *Suppose that $q \geq k$ and that Assumptions (1) and (4) hold. Moreover, assume that for all $Z = (Y, X) \in \mathcal{Z}$ with correctly specified parameter $\theta_0(F_Z)$ the map $A\colon \mathbb{R}^p \times \Theta \to \mathbb{R}^{q\times k}$ satisfies*

$$\mathbb{P}\big(\operatorname{rank}\big(A(X, \theta_0(F_Z))\big) = k\big) = 1. \tag{B.3}$$

If $\psi_A\colon \mathbb{R}\times\mathbb{R}^p\times\Theta \to \mathbb{R}^q$, $\psi_A(y, x, \theta) = A(x, \theta)\psi(y, x, \theta)$ is a strict unconditional $\mathcal{F}_\mathcal{Z}$-identification function for $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta$, then $\psi\colon \mathbb{R} \times \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ is a strict conditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0.

Example B.3 is again helpful because it shows that – in certain situations – we could relax the assumption that $A(X, \theta_0)$ needs to have full rank almost surely. However, relaxing the assumptions of Proposition B.5 does not seem to be a fruitful direction from our point of view, because one would need to tailor the relaxed assumptions almost on a case by case basis.

Recalling the discussion right after Definition 2.8, it is possible to derive a characterisation of (strict) identification functions for θ_0 in the spirit of Theorem B.1. For the sake of completeness, we state it explicitly.

Corollary B.6. *Let $\theta_0\colon \mathcal{F}_\mathcal{Z} \to \Theta \subseteq \mathbb{R}^q$ be a functional with a strict $\mathcal{F}_\mathcal{Z}$-identification function $\psi\colon \mathbb{R} \times \mathbb{R}^p \times \Theta \to \mathbb{R}^q$. Then the following two assertions hold:*

(i) *If $B\colon \Theta \to \mathbb{R}^{q\times q}$ is a matrix-valued function with $\det(B(\theta)) \neq 0$ for all $\theta \in \Theta$, then $(y, x, \theta) \mapsto B(\theta)\psi(y, x, \theta)$ is also a strict $\mathcal{F}_\mathcal{Z}$-identification function for θ_0.*

(ii) *Let $\psi'\colon \mathbb{R} \times \mathbb{R}^p \times \Theta \to \mathbb{R}^q$ be a strict $\mathcal{F}_\mathcal{Z}$-identification function for θ_0 such that ψ and ψ' both satisfy Assumption (5). Then there is a matrix-valued function $B\colon \operatorname{int}(\Theta) \to \mathbb{R}^{q\times q}$ with $\det(B(\theta)) \neq 0$ for all $\theta \in \Theta$ such that*

$$\bar{\psi}'(F_Z, \theta) = B(\theta)\bar{\psi}(F_Z, \theta)$$

for all $\theta \in \operatorname{int}(\Theta)$ and for all $F_Z \in \mathcal{F}_\mathcal{Z}$. If the integrated identification functions $\bar{\psi}(F_Z, \cdot)$ and $\bar{\psi}'(F_Z, \cdot)$ are continuous, then also B is continuous, which implies that either $\det(B(\theta)) > 0$ for all $\theta \in \Theta$ or $\det(B(\theta)) < 0$ for all $\theta \in \Theta$. Moreover, if $\mathcal{F}_\mathcal{Z}$ satisfies Assumption (6) and ψ, ψ' satisfy Assumption (7) it even holds that

$$\psi'(y, x, \theta) = B(\theta)\psi(y, x, \theta) \tag{B.4}$$

for almost all $(y, x, \theta) \in \mathbb{R} \times \mathbb{R}^p \times \operatorname{int}(\Theta)$.

The power of Corollary B.6 seems impressive. Roughly speaking, it says that starting with some strict $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-identification function $\varphi\colon \mathbb{R} \times \Xi \to \mathbb{R}^k$ for $\Gamma\colon \mathcal{F}_{\mathcal{Y}|\mathcal{X}} \to \Xi$ and after choosing some map $A\colon \mathbb{R}^p \times \Theta \to \mathbb{R}^{q\times k}$, one obtains the map $\mathbb{R} \times \mathbb{R}^p \times \Theta \ni (y, x, \theta) \mapsto \psi_A(y, x, \theta) := A(x, \theta)\varphi\big(y, m(x, \theta)\big) \in \mathbb{R}^q$. Then, one needs to check whether ψ_A is indeed an strict unconditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0. If one is lucky and it holds true, one knows that basically all other strict unconditional $\mathcal{F}_\mathcal{Z}$-identification functions for θ_0 must be of the form $B(\theta) \cdot A(X, \theta) \cdot \varphi\big(Y, m(X, \theta)\big)$, where $B(\theta)$ is a non-degenerate $q \times q$-matrix. However, the following example illustrates that one of the crucial conditions of Theorem B.1, namely the convexity of the class of distributions stipulated in Assumption (5), is often not satisfied in the framework of regression.

Example B.7 (Example B.3 continued)**.** Consider the same situation as in Example B.3 with $q = p = k = 1$. We saw that the choice $A(X, \theta) = X$ leads to a strict unconditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0, provided that $\mathbb{P}(X = 0) < 1$. However, using $\widetilde{A}(X, \theta) = X^3$ instead leads to $\mathbb{E}[\psi_{\widetilde{A}}(Y, X, \theta)] = \mathbb{E}[X^4](\theta - \theta_0)$, which is also a strict unconditional $\mathcal{F}_\mathcal{Z}$-identification function for θ_0. But if there are two marginal distributions in $\mathcal{F}_\mathcal{X}$ such that the ratios of second and forth moment $\mathbb{E}[X^2]/\mathbb{E}[X^4]$ are different, then there is no map $B(\theta)$ such that $\mathbb{E}[\psi_{\widetilde{A}}(Y,X,\theta)] = B(\theta)\mathbb{E}[\psi_A(Y,X,\theta)]$ for all $\theta \in \Theta$ and for all $X \in \mathcal{X}$. What has happened? Clearly, one of the assumptions of Corollary B.6 must be violated. And indeed, one can see that $\mathcal{F}_\mathcal{Z}$ defined at (B.2) is not convex implying that Assumption (5) is violated. To see this, let $X \in \mathcal{X}$. Then the distribution of $(X+\varepsilon, X)$ and (ε, X) is in $\mathcal{F}_\mathcal{Z}$. A convex mixture of the two distributions can be realised with the vector $(\eta X + \varepsilon, X)$ where $\mathbb{P}(\eta = 0) = 1 - \mathbb{P}(\eta = 1) \in (0,1)$ and η is independent of (X, ε). But the distribution of this vector is not contained in $\mathcal{F}_\mathcal{Z}$. □

C. Identification of the Efficient Z-estimator for double quantile models

Following Section 2, strict model consistency can directly be obtained by employing strictly consistent loss functions and a no-perfect collinearity condition of the model gradient. In contrast, this is more involved for the Z-estimator. Thus, the following proposition shows strict model identification for an efficient Z-estimator and for a large class of models.

Note that Theorem 3.1 asserts that the Z-estimator is efficient based on any choice $A^*_{t,C}(X_t,\theta)$ of instrument matrix such that $A^*_{t,C}(X_t,\theta_0) = CD_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$, see (3.7). This means we only have a condition on $A^*_{t,C}(X_t,\theta)$ for $\theta = \theta_0$. To come up with such a matrix, there are two straight forward ways how to guarantee this. First, we might set $A^*_{t,C}(X_t,\theta) = CD_t(X_t,\theta)^\intercal S_t(X_t,\theta)^{-1}$, and second, we might choose $A^*_{t,C}(X_t,\theta)$ to be constant in θ and equal to $CD_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$. For practical purposes, the latter situation is often hard or infeasible to implement, since it usually requires knowledge of the unknown true parameter θ_0 (and additional quantities of the conditional distribution F_t).

For the particular situation of the double quantile model, using the canonical identification function φ given in (4.4), $S_t(X_t,\theta_0)$ takes the form (4.5), which means it is entirely independent of any *any* knowledge on the underlying DGP whatsoever. This makes the latter choice attractive and reasonably feasible.

Proposition C.1. *We assume that (a) the double quantile model is linear with separated parameters, i.e.* $Q_\alpha(Y_t|X_t) = q_\alpha(X_t,\theta_0^\alpha) = X_t^\intercal\theta_0^\alpha$ *and* $Q_\beta(Y_t|X_t) = q_\alpha(X_t,\theta_0^\beta) = X_t^\intercal\theta_0^\beta$*, such that* $\theta_0 = (\theta_0^\alpha,\theta_0^\beta) \in \mathrm{int}(\Theta)$*, (b) for all* $t \in \mathbb{N}$*,* F_t *is differentiable with a strictly positive derivative* f_t*, and (c), there exists a possibly time-dependent deterministic constant* $c_t > 0$*, such that* $f_t(q_\alpha(X_t,\theta_0^\alpha)) = c_t f_t(q_\beta(X_t,\theta_0^\beta))$ *almost surely. Then, the moment function of the efficient Z-estimator of the DQR model is a strict* $\mathcal{F}_\mathcal{Z}$*-identification function for* θ_0*, i.e. it holds that*

$$\mathbb{E}\left[A^*_t(X_t,\theta_0)\varphi\big(Y_t, m(X_t,\theta)\big)\right] = 0 \quad \Longleftrightarrow \quad \theta = \theta_0,$$

where $A^*_t(X_t,\theta_0)$ *is given in* (3.7).

At the cost of some more tedious notation, Proposition C.1 can be generalised to the situation of linear models with not necessarily separated parameters, so long as there is at least one component that is used for modelling one quantile only, respectively. E.g. in a simple linear regression model, the two quantile models might have the same slope, but a different intercept, or vice versa, they might have the same intercept, but a different slope. Generalising the assertion much beyond linear models seems to be difficult due to the application of the mean value theorem in the proof.

D. The Two-Step Estimation Efficiency Bound

In related work, Barendse (2020) considers efficiency among the class of *two-step* estimators of semiparametric models for the quantile and ES with separated parameters. These two-step estimators utilise a quantile regression to estimate the quantile parameters in the first step and a restricted and weighted least squares estimator in the second step for the model parameters of the conditional ES. The author considers efficiency among the possible estimation weights from the second step weighted least squares estimator, see Barendse (2020) for details. This procedure amounts to efficiency of the ES parameters *in isolation*, which generally results in more restrictive efficiency bounds than efficiency of the joint model parameters considered in this article.

In our notation, the class of two-step estimators can be characterised by the general form (3.1), the identification functions in (4.15) and the class of instrument matrices[11]

$$A_t^{\dagger}(X_t, \theta_0) = \begin{pmatrix} \nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) & 0 \\ 0 & \phi_t''\big(e_\alpha(X_t, \theta_0^e)\big)\nabla_{\theta^e} e_\alpha(X_t, \theta_0^e) \end{pmatrix}. \tag{D.1}$$

Consequently, the family of two-step estimators of Barendse (2020) form a subclass of the general class of Z-estimators we consider in this article. Hence, it follows that the resulting two-step estimation efficiency bound is no smaller than the general Z-estimation efficiency bound of Theorem 3.1. While these two bounds can coincide in special situations, they generally do not as illustrated in the following.

For the special case of homogeneous location-scale models discussed in Section 4.3.2, the efficient weights of Barendse (2020) coincide with the choice of ϕ_t'' implied by a combination of (4.19) and (4.21) in Theorem 4.3. This illustrates that for this special case, and in terms of the ES parameters, θ^e, considered in isolation, the efficient two-step and the efficient M-estimator are equally efficient; see Barendse (2020, Section 4.4). However, if efficiency is considered for the full parameter vector, θ, the two-step estimator using the instrument matrix (D.1) is generally less efficient, which is caused by the inefficient choice of the first-step standard quantile regression. In this special case, joint efficiency could be guaranteed by employing an efficient quantile regression estimator in the first step, see e.g. Komunjer and Vuong (2010b,a). We refer to the simulation results of Section 5 and in particular to Table 1 for a numerical illustration.

More generally, Barendse (2020, Section 4.4) illustrates that, taken in isolation, the ES specific asymptotic sub-covariance matrix of the M-estimator $\widehat{\theta}^e$ is subject to his two-step efficiency bound. However, this does not hold if one considers the entire covariance matrix of the joint model parameters for the quantile and ES. This can be observed by comparing (D.1) with the efficient instrument matrix A_t^* given in (4.16) and (4.17): while $A_{t,C}^*$ generally requires non-zero off-diagonal blocks, the matrix $A_t^{\dagger}$ is restricted to a block diagonal matrix with zero off-diagonal blocks.

Recall that under the gradient condition that $\nabla_{\theta^q} q_\alpha(X_t, \theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t, \theta_0^e)$ $\forall t \in \mathbb{N}$ almost surely, part (A) of Theorem 4.3 implies that if conditions (4.19) or (4.20) fail to hold, the M-estimator cannot attain the Z-estimation efficiency bound. As Barendse (2020, Section 4.4) informally shows that the two-step estimators are equivalent to the class of M-estimators in terms of the efficiency of the ES parameters, this illustrates that the two-step estimators also cannot attain the Z-estimation efficiency bound in this setting. (Formally, relating $A_t^{\dagger}(X_t, \theta_0)$ in (D.1) to the efficient choice $A_t^*(X_t, \theta_0)$ and employing Theorem 3.1 as in the proof of Theorem

[11]Notice that for these estimators, the theory of Prokhorov and Schmidt (2009), Bartalotti (2013) can be used to establish that the asymptotic distribution of the joint Z- and the two-step estimators coincide.

4.3 yields the desired result.) Besides supporting our claim of an existing *efficiency gap* for the joint quantile and ES models, this illustrates that the two-step efficiency bound of Barendse (2020) does generally not coincide with the general Z-estimation efficiency bound of Hansen (1985), Chamberlain (1987), and Newey (1993).

We illustrate the theoretical considerations of this section numerically through the simulation setup of Section 5. In Panel A of Table 1, we additionally report the two-step efficiency bound in the line denoted "Barendse Bound". For the homoskedastic innovations and for the ES specific parameters, the two-step efficiency bound coincides with the Z-estimation efficiency bound, while it does not for the quantile parameters. This is primarily caused by the inefficient first-step quantile estimation – using an efficient quantile estimator (based on $g_t(\xi_1) = F_t(\xi_1)$) would equate both efficiency bounds in the homoskedastic case. In contrast, in the heteroskedastic case, the two-step efficiency bound is considerably larger than the Z-estimation efficiency bound for all four considered parameters. Interestingly, the choice of ϕ^{eff2} motivated by this two-step estimation efficiency bound exhibits equally efficient ES parameters while the quantile parameters show larger standard deviations.

E. Additional Simulation Results

In this section, we report simulation results for the double quantile model discussed in Section 4.2. (Only Table 5 reports results for additional probability levels for the joint quantile and ES models, discussed in Section 5.)

For the double quantile models, we simulate data according to the process given in (4.11), where we employ linear models for the conditional location and scale,

$$X_t \overset{iid}{\sim} 3\,\text{Beta}(3, 1.5), \quad \mu(X_t, \eta_0) = 10 + 0.5X_t, \quad \text{and} \quad \sigma(X_t, \eta_0) = 0.5 + 0.5X_t.$$

For the model innovations u_t, we choose the following three different specifications:

(a) $u_t \overset{iid}{\sim} \mathcal{N}(0, 1)$;

(b) $u_t \sim t_{\nu_t}(\mu_t, \sigma_t^2)$ with time-varying degrees of freedom, $\nu_t = 3\,\mathbb{1}_{\{t \le T/2\}} + 100\,\mathbb{1}_{\{t > T/2\}}$, where μ_t and σ_t are given in (4.13);

(c) $u_t \sim \mathcal{SN}(\mu_t, \sigma_t^2, \gamma_t)$ follows a skewed normal distribution with time-varying skewness, $\gamma_t = 0.9\,\mathbb{1}_{\{t > T/2\}}$, where μ_t and σ_t are given in (4.14).

These choices are motivated through the theoretical considerations of Section 4.2 that for location-scale models with i.i.d. residuals, the M-estimator is able to attain the Z-estimation efficiency bound, while conversely, it cannot do so for heterogeneously distributed innovations. The heterogeneously skewed process given in (c) is motivated by symmetric prediction intervals where $\alpha = 1 - \beta$. For the considered processes, we estimate linear quantile models with *separated* model parameters,

$$q_\alpha(X_t, \theta) = \theta^{(1)} + \theta^{(2)} X_t, \quad \text{and} \quad q_\beta(X_t, \theta) = \theta^{(3)} + \theta^{(4)} X_t, \tag{E.1}$$

In order to consider models with joint parameters, we use a slightly modified parameterization of the process by using $\sigma(X_t, \eta_0) = 0.5X_t$, which implies that $Q_\alpha(Y_t|X_t) = 10 + (0.5 + 0.5z_\alpha)X_t$, and $Q_\beta(Y_t|X_t) = 10 + (0.5 + 0.5z_\beta)X_t$. Hence, we use the (correctly specified) *joint intercept models*

$$q_\alpha(X_t, \theta) = \theta^{(1)} + \theta^{(2)} X_t, \quad \text{and} \quad q_\beta(X_t, \theta) = \theta^{(1)} + \theta^{(3)} X_t. \tag{E.2}$$

Table 2.: Asymptotic Standard Deviations of Double Quantile Models

	(a) Homoskedastic				(b) Heteroskedastic t				(c) Heteroskedastic $\mathcal{SN}$			
$g_t(\xi)$	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
	Panel A: Separated Model Parameters and $(\alpha, \beta) = (1\%, 2.5\%)$											
ξ	14.220	7.865	10.175	5.627	51.446	24.383	31.131	15.319	11.976	5.552	9.904	4.825
$\exp(\xi)$	13.760	7.663	9.758	5.410	83.311	47.176	36.946	19.201	14.020	6.644	10.522	5.149
$\log(\xi)$	14.480	8.003	10.298	5.693	52.772	25.134	31.526	15.533	11.884	5.511	9.900	4.825
$F_{\text{Log}}(\xi)$	14.219	7.864	10.175	5.627	51.473	24.396	31.129	15.317	11.976	5.552	9.904	4.825
$F_t(\xi)$	13.619	7.546	9.745	5.399	48.632	22.643	29.936	14.636	10.808	4.941	9.368	4.526
Eff. B.	13.619	7.546	9.745	5.399	47.871	22.289	29.466	14.409	10.686	4.881	9.259	4.471
	Panel B: Separated Model Parameters and $(\alpha, \beta) = (5\%, 95\%)$											
ξ	8.006	4.449	8.006	4.449	14.663	8.017	14.663	8.017	5.249	3.216	9.860	5.307
$\exp(\xi)$	7.704	4.285	14.095	7.181	14.188	7.813	31.497	15.287	4.860	3.001	16.512	8.278
$\log(\xi)$	8.066	4.481	7.867	4.375	14.941	8.162	14.363	7.859	5.311	3.250	9.725	5.236
$F_{\text{Log}}(\xi)$	8.006	4.449	8.006	4.449	14.663	8.016	14.663	8.017	5.249	3.216	9.860	5.307
$F_t(\xi)$	7.672	4.271	7.672	4.271	13.706	7.507	13.706	7.507	4.319	2.751	9.394	5.077
Eff. B.	7.672	4.271	7.672	4.271	13.706	7.507	13.706	7.507	4.317	2.750	9.389	5.075
	Panel C: Separated Model Parameters and $(\alpha, \beta) = (1\%, 90\%)$											
ξ	14.251	7.899	6.525	3.617	37.583	22.319	11.207	6.031	8.884	6.205	8.429	4.525
$\exp(\xi)$	13.792	7.698	10.355	5.368	40.573	27.267	18.742	9.397	7.000	5.135	13.009	6.599
$\log(\xi)$	14.512	8.038	6.421	3.561	43.160	25.308	11.057	5.951	9.276	6.416	8.317	4.466
$F_{\text{Log}}(\xi)$	14.250	7.899	6.525	3.617	37.370	22.182	11.207	6.031	8.883	6.204	8.429	4.525
$F_t(\xi)$	13.650	7.579	6.250	3.471	31.101	19.014	10.886	5.856	5.558	4.488	7.896	4.262
Eff. B.	13.650	7.579	6.250	3.471	31.097	19.012	10.885	5.856	5.556	4.487	7.893	4.261
	Panel D: Joint Model Parameters and $(\alpha, \beta) = (1\%, 2.5\%)$											
ξ	4.996	3.519	3.167		25.108	13.458	12.548		7.838	3.750	3.752	
$\exp(\xi)$	4.214	3.058	2.770		22.224	14.042	12.152		7.362	3.617	3.599	
$\log(\xi)$	5.130	3.598	3.234		26.547	14.234	13.162		7.932	3.791	3.791	
$F_{\text{Log}}(\xi)$	4.996	3.519	3.167		25.102	13.453	12.546		7.838	3.750	3.752	
$F_t(\xi)$	3.768	2.885	2.569		15.344	9.951	8.759		5.007	2.704	2.687	
Eff. B.	3.757	2.880	2.564		14.649	9.672	8.459		4.991	2.690	2.673	
	Panel E: Joint Model Parameters and $(\alpha, \beta) = (5\%, 95\%)$											
ξ	2.825	1.961	1.961		5.003	3.496	3.496		2.092	1.495	1.799	
$\exp(\xi)$	7.551	4.158	4.182		17.954	10.555	9.342		8.000	4.636	4.517	
$\log(\xi)$	2.807	1.974	1.928		5.040	3.572	3.450		2.027	1.481	1.746	
$F_{\text{Log}}(\xi)$	2.825	1.961	1.961		5.003	3.496	3.496		2.092	1.495	1.799	
$F_t(\xi)$	2.173	1.648	1.648		3.841	2.903	2.903		1.341	1.099	1.466	
Eff. B.	2.173	1.648	1.648		3.841	2.903	2.903		1.339	1.098	1.465	
	Panel F: Joint Model Parameters and $(\alpha, \beta) = (1\%, 90\%)$											
ξ	2.882	2.539	1.892		5.358	6.276	3.265		2.695	2.261	1.911	
$\exp(\xi)$	5.890	3.871	3.287		12.618	14.475	6.535		7.405	5.533	4.061	
$\log(\xi)$	2.796	2.554	1.840		5.364	6.789	3.231		2.530	2.235	1.820	
$F_{\text{Log}}(\xi)$	2.882	2.539	1.892		5.357	6.272	3.265		2.695	2.261	1.911	
$F_t(\xi)$	2.419	2.343	1.650		3.629	6.370	2.467		1.805	2.010	1.463	
Eff. B	2.272	2.292	1.578		2.301	4.575	1.960		1.102	1.310	1.225	

This table presents the (approximated) asymptotic standard deviations for semiparametric double quantile models at different probability levels in the horizontal panels. Panels A-C report results for the models with separated parameters given in (E.1) while Panels D-F considers the joint intercept models given in (E.2). Results for the three residual distributions described in Section E are reported in the three vertical panels of the table. We furthermore consider four classical choices of $g_t(\xi)$ together with the (pseudo-) efficient choice $F_t(\xi)$ and the Z-estimation efficiency bound.

Table 2 presents the approximated standard deviations of the estimated parameters of the joint quantile models for the probability levels $(\alpha, \beta) \in \{(1\%, 2.5\%), (5\%, 95\%), (1\%, 90\%)\}$, where the first choice is important for VaR modeling in financial risk management, while the remaining two consider estimation of a symmetric and an asymmetric prediction interval. Table 3 and Table 4 show results for additional pairs of probability levels. Panels A-C consider the separated parameter models in (E.1) while Panels D-F consider models with joint parameters in (E.2). We show results for the joint M-estimator using the general loss function in (4.6) paired with the choices $g_t(\xi) = g_{1,t}(\xi) = g_{2,t}(\xi)$ given in the first column of Table 2 together with the Z-estimation Efficiency Bound, where F_{Log} denotes the distribution function of a standard

Table 3.: Asy. Standard Deviations of Separated Parameter Double Quantile Models

	(a) Homoskedastic				(b) Heteroskedastic t				(c) Heteroskedastic $\mathcal{SN}$			
$g_t(\xi)$	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
	Panel A: $(\alpha, \beta) = (0.5\%, 1\%)$											
ξ	19.103	10.518	14.620	8.050	85.963	39.108	59.937	27.987	14.020	6.283	12.058	5.577
$\exp(\xi)$	18.753	10.449	14.170	7.853	224.734	139.513	106.616	59.495	17.877	8.329	14.088	6.662
$\log(\xi)$	19.539	10.749	14.883	8.189	4914.649	2373.196	60.457	28.334	13.805	6.185	11.972	5.539
$F_{log}(\xi)$	19.101	10.517	14.619	8.049	88.549	40.686	60.019	28.032	14.021	6.284	12.059	5.577
$F_t(\xi)$	18.321	10.104	14.022	7.733	76.920	34.320	55.830	25.659	11.999	5.297	10.859	4.953
Eff. B.	18.321	10.104	14.022	7.733	75.811	33.848	55.029	25.315	11.893	5.249	10.762	4.909
	Panel B: $(\alpha, \beta) = (5\%, 10\%)$											
ξ	8.238	4.517	6.664	3.654	17.012	8.929	11.996	6.389	7.871	4.009	7.053	3.728
$\exp(\xi)$	7.926	4.351	6.515	3.575	17.276	9.145	11.719	6.243	7.904	4.020	6.966	3.680
$\log(\xi)$	8.299	4.550	6.684	3.665	17.223	9.040	12.064	6.425	7.892	4.021	7.067	3.736
$F_{log}(\xi)$	8.237	4.517	6.664	3.654	17.012	8.929	11.996	6.389	7.871	4.009	7.053	3.728
$F_t(\xi)$	7.895	4.337	6.387	3.508	16.605	8.672	11.643	6.208	7.647	3.871	6.858	3.626
Eff. B.	7.895	4.337	6.387	3.508	16.453	8.585	11.536	6.145	7.569	3.821	6.784	3.577
	Panel C: $(\alpha, \beta) = (25\%, 50\%)$											
ξ	5.306	2.910	4.880	2.677	6.306	3.450	5.436	2.975	4.734	2.604	5.261	2.916
$\exp(\xi)$	5.500	3.011	5.646	3.054	6.470	3.536	6.312	3.406	4.904	2.692	6.101	3.332
$\log(\xi)$	5.290	2.902	4.840	2.656	6.291	3.443	5.390	2.951	4.720	2.597	5.216	2.893
$F_{log}(\xi)$	5.306	2.910	4.880	2.677	6.306	3.450	5.436	2.975	4.734	2.604	5.261	2.916
$F_t(\xi)$	5.091	2.797	4.683	2.573	6.037	3.309	5.211	2.857	4.530	2.496	5.016	2.784
Eff. B.	5.091	2.797	4.683	2.573	6.031	3.306	5.206	2.854	4.491	2.473	4.973	2.759
	Panel D: $(\alpha, \beta) = (1\%, 99\%)$											
ξ	14.506	7.985	14.506	7.985	40.113	22.072	40.113	22.072	7.883	7.191	17.077	8.910
$\exp(\xi)$	14.062	7.792	30.706	15.013	48.582	30.747	153.517	68.048	5.655	5.813	32.722	15.639
$\log(\xi)$	14.765	8.122	14.221	7.835	44.642	24.482	39.071	21.519	8.317	7.442	16.911	8.822
$F_{log}(\xi)$	14.505	7.985	14.506	7.985	39.958	21.974	40.113	22.072	7.881	7.190	17.077	8.910
$F_t(\xi)$	13.918	7.674	13.918	7.674	35.871	19.775	35.871	19.775	3.799	4.523	16.317	8.576
Eff. B.	13.918	7.674	13.918	7.674	35.871	19.775	35.871	19.775	3.799	4.523	16.316	8.576
	Panel E: $(\alpha, \beta) = (10\%, 99\%)$											
ξ	6.780	3.710	14.807	8.102	11.300	6.049	36.902	21.901	4.317	2.493	17.055	9.011
$\exp(\xi)$	6.627	3.628	30.733	15.003	11.007	5.893	171.522	76.831	4.179	2.420	33.773	16.214
$\log(\xi)$	6.800	3.721	14.507	7.944	11.368	6.085	35.250	21.028	4.335	2.503	16.804	8.879
$F_{log}(\xi)$	6.780	3.710	14.807	8.102	11.300	6.049	36.902	21.901	4.317	2.493	17.055	9.011
$F_t(\xi)$	6.493	3.559	14.180	7.772	10.979	5.875	29.813	18.301	3.714	2.186	16.413	8.700
Eff. B.	6.493	3.559	14.180	7.772	10.978	5.874	29.809	18.299	3.713	2.186	16.410	8.699

This table presents the (approximated) asymptotic standard deviations for semiparametric double quantile models with separated model parameters given in (E.1) at different probability levels in the horizontal panels. Results for the three residual distributions described in Section E are reported in the three vertical panels of the table. We furthermore consider four classical choices of $g_t(\xi)$ together with the (pseudo-) efficient choice $F_t(\xi)$ and the Z-estimation efficiency bound.

logistic distribution. We approximate the true asymptotic covariance through 1000 simulation replications each consisting of a sample size of $T = 2000$.

The numerical results generally confirm the conclusions of Section 4.2: the pseudo-efficient M-estimator with $g_t(\xi) = F_t(\xi)$ attains the efficiency bound for location-scale processes with homoskedastic innovation distributions, while it generally cannot attain the efficiency bound for both heteroskedastic processes. Furthermore, as discussed in Section 4.2, for symmetric quantile levels as in Panel B, the symmetrically heteroskedastic process in (b) is not sufficient for generating an efficiency gap, whereas the heteroskedastic process in (c) is sufficient. Remarkably, even for models with separated parameters, where the pseudo-efficient choices correspond to efficient estimators for both quantile models individually (Komunjer and Vuong, 2010b,a), the joint M-estimator based on these choices does not attain the efficiency bound for the processes with heteroskedastic innovations. Furthermore, the first four choices of loss functions result in an anticipated loss of efficiency, but are straight-forward to implement as they do not depend on any unknown quantities.

We observe that in these situations, the gap becomes numerically larger for quantile levels

Table 4.: Asy. Standard Deviations of Joint Parameter Double Quantile Models

	(a) Homoskedastic			(b) Heteroskedastic t			(c) Heteroskedastic $\mathcal{SN}$		
$g_t(\xi)$	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
	Panel A: $(\alpha, \beta) = (0.5\%, 1\%)$								
ξ	7.022	4.842	4.459	53.275	25.708	25.041	10.381	4.710	4.720
$\exp(\xi)$	5.766	4.122	3.821	55.213	37.705	33.190	9.686	4.600	4.599
$\log(\xi)$	7.273	4.985	4.586	57.732	27.990	26.848	10.543	4.774	4.781
$F_{\text{Log}}(\xi)$	7.022	4.842	4.459	53.229	25.674	25.024	10.381	4.710	4.720
$F_t(\xi)$	5.258	3.922	3.591	32.705	18.216	17.152	6.537	3.287	3.274
Eff. Bound	5.216	3.903	3.570	30.869	17.564	16.438	6.491	3.267	3.254
	Panel B: $(\alpha, \beta) = (5\%, 10\%)$								
ξ	3.127	2.136	2.003	7.937	4.737	4.417	4.885	2.566	2.581
$\exp(\xi)$	2.880	1.984	1.882	7.194	4.293	4.068	4.703	2.475	2.497
$\log(\xi)$	3.158	2.155	2.018	8.076	4.820	4.481	4.910	2.580	2.593
$F_{\text{Log}}(\xi)$	3.126	2.136	2.003	7.937	4.737	4.417	4.885	2.566	2.581
$F_t(\xi)$	2.365	1.740	1.625	5.638	3.736	3.423	3.351	1.965	1.965
Eff. Bound	2.360	1.737	1.622	5.417	3.637	3.318	3.338	1.952	1.952
	Panel C: $(\alpha, \beta) = (25\%, 50\%)$								
ξ	2.087	1.404	1.370	2.344	1.610	1.535	2.324	1.445	1.512
$\exp(\xi)$	2.534	1.594	1.613	2.833	1.812	1.802	2.787	1.645	1.768
$\log(\xi)$	2.060	1.392	1.354	2.317	1.598	1.518	2.298	1.433	1.496
$F_{\text{Log}}(\xi)$	2.087	1.404	1.370	2.344	1.610	1.535	2.324	1.445	1.512
$F_t(\xi)$	1.630	1.172	1.141	1.759	1.312	1.246	1.925	1.245	1.304
Eff. Bound	1.630	1.172	1.141	1.757	1.310	1.244	1.883	1.222	1.280
	Panel D: $(\alpha, \beta) = (1\%, 99\%)$								
ξ	4.866	3.409	3.409	13.386	9.432	9.432	2.982	2.650	2.835
$\exp(\xi)$	17.887	10.417	9.267	102.533	80.324	46.348	19.564	14.581	9.957
$\log(\xi)$	4.902	3.485	3.363	14.724	10.450	9.711	3.018	2.719	2.800
$F_{\text{Log}}(\xi)$	4.866	3.409	3.409	13.380	9.427	9.431	2.982	2.650	2.835
$F_t(\xi)$	3.728	2.864	2.864	9.913	7.560	7.560	1.310	1.269	2.329
Eff. Bound	3.728	2.864	2.864	9.913	7.560	7.560	1.310	1.269	2.329
	Panel E: $(\alpha, \beta) = (10\%, 99\%)$								
ξ	2.878	1.890	2.536	5.297	3.249	6.249	1.904	1.286	2.471
$\exp(\xi)$	11.803	6.142	6.857	78.737	40.679	38.083	10.068	5.367	6.326
$\log(\xi)$	2.898	1.903	2.506	5.424	3.319	6.051	1.899	1.287	2.436
$F_{\text{Log}}(\xi)$	2.878	1.890	2.536	5.297	3.249	6.249	1.904	1.286	2.471
$F_t(\xi)$	2.409	1.644	2.339	3.474	2.415	6.176	1.453	1.027	2.287
Eff. Bound	2.262	1.572	2.288	2.219	1.941	4.497	1.433	1.016	2.282

This table presents the (approximated) asymptotic standard deviations for semiparametric double quantile models with joint model parameters given in (E.2) at different probability levels in the horizontal panels. Results for the three residual distributions described in Section E are reported in the three vertical panels of the table. We furthermore consider four classical choices of $g_t(\xi)$ together with the (pseudo-) efficient choice $F_t(\xi)$ and the Z-estimation efficiency bound.

Table 5.: Asy. Standard Deviations of Separated Parameter Quantile and ES Models

		(a) Homoskedastic				(b) Heteroskedastic			
$g_t(\xi_1)$	$\phi_t(\xi_2)$	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
		Panel A: $\alpha = 1\%$ and Models with Separated Parameters							
0	$\exp(\xi_2)$	17.892	11.444	19.444	11.586	250.078	224.646	315.156	292.206
$F_t(\xi_1)$	$\exp(\xi_2)$	17.621	11.101	19.444	11.586	144.249	68.741	315.156	292.206
0	$F_{\text{Log}}(\xi_2)$	17.731	11.301	19.221	11.402	249.685	224.195	314.474	291.337
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	17.465	10.964	19.221	11.402	144.007	68.615	314.474	291.337
0	$-\log(-\xi_2)$	13.887	7.653	17.098	9.422	70.473	32.802	153.598	84.660
$F_t(\xi_1)$	$-\log(-\xi_2)$	13.887	7.653	17.098	9.422	70.463	32.796	153.598	84.660
0	$\phi_t^{\text{eff1}}(\xi_2)$	13.879	7.649	17.058	9.401	70.863	33.001	153.510	84.614
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	13.879	7.649	17.058	9.401	70.556	32.821	153.510	84.614
0	$\phi_t^{\text{eff2}}(\xi_2)$	13.879	7.649	17.058	9.401	73.239	33.737	144.306	79.537
Barendse Bound		14.476	7.964	17.058	9.401	66.577	30.839	144.306	79.537
Efficiency Bound		13.879	7.649	17.058	9.401	55.070	24.918	123.549	70.916
		Panel B: $\alpha = 10\%$ and Models with Separated Parameters							
0	$\exp(\xi_2)$	7.263	4.290	7.477	4.224	19.110	12.139	28.101	17.067
$F_t(\xi_1)$	$\exp(\xi_2)$	6.690	3.734	7.477	4.224	14.568	7.785	28.101	17.067
0	$F_{\text{Log}}(\xi_2)$	7.183	4.225	7.405	4.167	18.927	11.983	27.787	16.803
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	6.647	3.705	7.405	4.167	14.478	7.735	27.787	16.803
0	$-\log(-\xi_2)$	6.359	3.503	7.185	3.957	13.351	7.048	24.002	13.226
$F_t(\xi_1)$	$-\log(-\xi_2)$	6.359	3.503	7.185	3.957	13.320	7.035	24.002	13.226
0	$\phi_t^{\text{eff1}}(\xi_2)$	6.353	3.500	7.157	3.943	13.385	7.065	23.952	13.199
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	6.353	3.500	7.157	3.943	13.296	7.028	23.952	13.199
0	$\phi_t^{\text{eff2}}(\xi_2)$	6.353	3.500	7.157	3.943	13.943	7.315	22.464	12.378
Barendse Bound		6.628	3.645	7.157	3.943	13.591	7.176	22.464	12.378
Efficiency Bound		6.353	3.500	7.157	3.943	12.629	6.742	20.111	11.274
		Panel C: $\alpha = 1\%$ and Models with Joint Parameters							
0	$\exp(\xi_2)$	5.578	4.175	4.257		65.432	73.420	71.061	
$F_t(\xi_1)$	$\exp(\xi_2)$	5.563	4.146	4.246		62.764	63.979	67.978	
0	$F_{\text{Log}}(\xi_2)$	5.573	4.116	4.218		64.370	70.814	68.972	
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	5.558	4.089	4.208		61.655	61.461	65.878	
0	$-\log(-\xi_2)$	5.611	3.767	4.002		43.561	22.540	31.215	
$F_t(\xi_1)$	$-\log(-\xi_2)$	5.609	3.766	4.001		43.548	22.535	31.209	
0	$\phi_t^{\text{eff1}}(\xi_2)$	5.423	3.676	3.912		45.628	24.529	32.981	
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	5.420	3.675	3.911		45.452	24.440	32.900	
0	$\phi_t^{\text{eff2}}(\xi_2)$	5.303	3.616	3.856		46.491	24.109	32.104	
Efficiency Bound		5.153	3.542	3.786		25.453	15.221	23.534	
		Panel D: $\alpha = 10\%$ and Models with Joint Parameters							
0	$\exp(\xi_2)$	2.452	1.711	1.748		7.974	5.334	5.898	
$F_t(\xi_1)$	$\exp(\xi_2)$	2.387	1.645	1.712		6.984	4.322	5.341	
0	$F_{\text{Log}}(\xi_2)$	2.471	1.707	1.750		7.964	5.246	5.834	
$F_t(\xi_1)$	$F_{\text{Log}}(\xi_2)$	2.392	1.643	1.709		6.897	4.246	5.258	
0	$-\log(-\xi_2)$	2.596	1.739	1.791		7.411	4.274	5.278	
$F_t(\xi_1)$	$-\log(-\xi_2)$	2.545	1.715	1.768		7.211	4.192	5.199	
0	$\phi_t^{\text{eff1}}(\xi_2)$	2.362	1.624	1.684		7.388	4.367	5.326	
$F_t(\xi_1)$	$\phi_t^{\text{eff1}}(\xi_2)$	2.355	1.620	1.681		7.202	4.283	5.242	
0	$\phi_t^{\text{eff2}}(\xi_2)$	2.331	1.609	1.669		7.527	4.378	5.231	
Efficiency Bound		2.310	1.598	1.659		4.515	3.108	3.929	

This table presents the (approximated) asymptotic standard deviations for semiparametric joint quantile and ES models at joint probability level of 1% and 10% for various choices of M-estimators together with the Z-estimation efficiency bound and in Panel A and B, the two-step efficiency bound of Barendse (2020). Panel A and B report results for the models with separated parameters given in (5.2) while Panel C and D consider the joint intercept models given in (5.3). The two considered residual distributions are presented in the two vertical panels of the table.

in the tails of the conditional distributions (Panel A) and for quantile levels which are close together. This makes it particularly relevant in the VaR literature, where VaR is often reported for multiple and extreme quantile levels. The first observation can be explained by condition (4.7), i.e. for a numerically large efficiency gap, one requires heterogeneity of the conditional density functions at the respective quantile levels $f_t(q_\alpha(X_t, \theta_0^\alpha))$ and $f_t(q_\beta(X_t, \theta_0^\beta))$. Such a heterogeneity is generally easier to achieve in the tails of the conditional distributions rather than in their central regions. The second observation above can be explained by noting that the efficiency gap is basically driven by the non-zero term $\alpha(1-\beta)$ in the off-diagonal entries of the matrix $S_t(X_t, \theta_0)$ in (4.5). As $\alpha < \beta$ by assumption, this term is particularly large for $\alpha \approx \beta \approx 1/2$ and particularly small for $\alpha \ll \beta$.

Panels D-F in Table 2 present results for the models with a joint intercept parameter. We find that the general Z-estimation efficiency bound is still valid, which substantiates the statement of Theorem 3.1. Differently from models with separated parameters, the pseudo-efficient choices $g_t(\xi) = F_t(\xi)$ generally cannot attain the efficiency bound, even in the homoskedastic residual case, which indicates that the efficiency gap applies to an even wider class of processes. For both heteroskedastic innovation distributions, the efficiency gap exists and is larger in magnitude. Furthermore, the efficiency gap becomes substantially larger, especially in the example of Panel F, while the pseudo-efficient choices still result in the most efficient estimator among the considered choices. These results show that the efficiency gap is present for a large class of double quantile models and data generating processes, which goes beyond the theoretically considered models of Theorem 4.2.

F. Proofs

Proof of Theorem 2.5. (i) Let $Z = (Y, X) \in \mathcal{Z}$ and $\theta_0 = \theta_0(F_Z)$. Suppose that $\theta \neq \theta_0$. Then $\mathbb{E}\big[\rho\big(Y, m(X, \theta_0)\big)\big|X\big] = \mathbb{E}\big[\rho\big(Y, \Gamma(F_{Y|X})\big)\big|X\big] \le \mathbb{E}\big[\rho\big(Y, m(X, \theta)\big)\big|X\big]$ due to the $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistency of ρ. Invoking Assumption (1), we get $\mathbb{P}(m(X, \theta_0) \neq m(X, \theta)) > 0$. Therefore, if ρ is strictly $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$-consistent for Γ, it is also strictly $\mathcal{F}_{\mathcal{Z}}$-model-consistent for m.

(ii) Suppose that ρ is conditionally $\mathcal{F}_{\mathcal{Z}}$-model consistent and let $(Y, X) \in \mathcal{Z}$. Then for all $\theta \in \Theta$ it holds that $\mathbb{P}\Big(\mathbb{E}\big[\rho\big(Y, m(X, \theta_0)\big)\big|X\big] \le \mathbb{E}\big[\rho\big(Y, m(X, \theta)\big)\big|X\big]\Big) = 1$. Since the countable union of null sets is again a null set, this implies that

$$\mathbb{P}\Big(\mathbb{E}\big[\rho\big(Y, m(X, \theta_0)\big)\big|X\big] \le \mathbb{E}\big[\rho\big(Y, m(X, \theta)\big)\big|X\big] \; \forall \theta \in \Theta \cap \mathbb{Q}^q\Big) = 1,$$

where $\mathbb{Q}$ is the set of all rationals. Using the fact that $\Theta \cap \mathbb{Q}^q$ is dense in Θ and due to the stipulated continuity of the conditional expectations of the losses in Assumption (3), we obtain that $\mathbb{P}(A) = 1$ with

$$A = \Big\{\omega \in \Omega \,|\, \mathbb{E}\big[\rho\big(Y, m(X, \theta_0)\big)\big|X\big](\omega) = \mathbb{E}\big[\rho\big(Y, \Gamma(F_{Y|X})\big)\big|X\big](\omega) \le \mathbb{E}\big[\rho\big(Y, m(X, \theta)\big)\big|X\big](\omega) \; \forall \theta \in \Theta\Big\} \in \mathcal{A},$$

where we also used Assumption (1). Let $A' \in \mathcal{A}$ be the set with probability one such that $m(X(\omega), \cdot)$ is surjective for all $\omega \in A'$. Then for all $\omega \in A \cap A'$ we get that

$$\int \rho\big(y, \Gamma(F_{Y|X}(\cdot, \omega))\big) F_{Y|X}(\mathrm{d}y, \omega) \le \int \rho\big(y, m(X(\omega), \theta)\big) F_{Y|X}(\mathrm{d}y, \omega)$$

for all $\theta \in \Theta$. Finally, exploiting the surjectivity of the model, we arrive at the claim. Clearly, it is only possible to establish this assertion on a version of $\mathcal{F}_{\mathcal{Y}|\mathcal{X}}$; see footnote 1.

(iii) This is a standard application of the tower property together with the positivity of the expectation.

(iv) Assume that ρ is not strictly conditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent for m. That means there exists $Z = (Y, X) \in \mathcal{Z}$ with true parameter $\theta_0 = \theta_0(F_Z)$ such that for some $\theta \neq \theta_0$ the event $A = \big\{\omega \,|\, \mathbb{E}\big[\rho\big(Y, m(X,\theta)\big) - \rho\big(Y, m(X,\theta_0)\big)\big|X\big](\omega) \le 0\big\}$ has positive probability. Let $\widetilde{Z} = (\widetilde{Y}, \widetilde{X}) \in \mathcal{Z}$ be the pair given by Assumption (4) with A specified above. Then clearly

$$\mathbb{E}\big[\rho\big(\widetilde{Y}, m(\widetilde{X},\theta)\big) - \rho\big(\widetilde{Y}, m(\widetilde{X},\theta_0)\big)\big] = \mathbb{E}\Big[\mathbb{E}\big[\rho\big(\widetilde{Y}, m(\widetilde{X},\theta)\big) - \rho\big(\widetilde{Y}, m(\widetilde{X},\theta_0)\big)\big|\widetilde{X}\big]\Big] \le 0. \quad \text{(F.1)}$$

This means that ρ is not strictly unconditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent for m. The argument when we assume that ρ is merely conditionally $\mathcal{F}_{\mathcal{Z}}$-model-consistent works analogously, where we replace the inequalities in the definition of A and in (F.1) with strict inequalities. ☐

Proof of Lemma 2.9. Part (i) is a direct application of the definitions, using similar arguments to the ones in the proof of Theorem 2.5 (i). For part (ii) we have under Assumption (1) that $0 = \mathbb{E}\big[\varphi\big(Y, m(X,\theta_0)\big)\big|X\big] = \mathbb{E}\big[\varphi\big(Y, \Gamma(F_{Y|X})\big)\big|X\big]$. ☐

Proof of Theorem 3.1. For $A^*_{t,C}(X_t,\theta_0) = CD_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$, one obtains that $\Delta_{T,\mathbb{A}^*} = \frac{1}{T}\sum_{t=1}^T C\,\mathbb{E}\left[D_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1} D_t(X_t,\theta_0)\right]$ and $\Sigma_{T,\mathbb{A}^*} = \Delta_{T,\mathbb{A}^*}C^\intercal$. Thus, the asymptotic covariance of the Z-estimator based on the choice $A^*_{t,C}(X,\theta_0)$ has the asymptotic covariance $\Delta^{-1}_{T,\mathbb{A}^*}\Sigma_{T,\mathbb{A}^*}\left(\Delta^{-1}_{T,\mathbb{A}^*}\right)^\intercal = \Lambda_T^{-1} = \left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[D_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1} D_t(X_t,\theta_0)\right]\right)^{-1}$ for all deterministic and non-singular choices of C, which shows part (i) of Theorem 3.1.

As the asymptotic covariance is independent of the choice of C, without loss of generality we continue with $C = I_q$ for the proof of part (ii) and henceforth use the notation $A^*_t = A^*_{t,I_q}$. In order to show that the matrix $\Delta^{-1}_{T,\mathbb{A}}\Sigma_{T,\mathbb{A}}(\Delta^{-1}_{T,\mathbb{A}})^\intercal - \Lambda_T^{-1}$ is positive semi-definite for all $T \ge 1$ and for all instrument matrices $A_t(X_t,\theta)$, we define the random vector

$$\chi_{t,T} = \Big(\Delta^{-1}_{T,\mathbb{A}}A_t(X_t,\theta_0) - \Lambda_T^{-1}A^*_t(X_t,\theta_0)\Big)\,\varphi\big(Y_t, m(X_t,\theta_0)\big),$$

for all $t, 1 \le t \le T, T \ge 1$. Then $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\chi_{t,T}\chi_{t,T}^\intercal\right]$ equals

$$\begin{aligned}
&\Delta^{-1}_{T,\mathbb{A}}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A_t(X_t,\theta_0)\varphi(Y_t,m(X_t,\theta_0))\varphi(Y_t,m(X_t,\theta_0))^\intercal A_t(X_t,\theta_0)^\intercal\right]\right)(\Delta^\intercal_{T,\mathbb{A}})^{-1}\\
&+\Lambda_T^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A^*_t(X_t,\theta_0)\varphi(Y_t,m(X_t,\theta_0))\varphi(Y_t,m(X_t,\theta_0))^\intercal A^*_t(X_t,\theta_0)^\intercal\right]\right)(\Lambda_T^{-1})^\intercal\\
&-\Delta^{-1}_{T,\mathbb{A}}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A_t(X_t,\theta_0)\varphi(Y_t,m(X_t,\theta_0))\varphi(Y_t,m(X_t,\theta_0))^\intercal A^*_t(X_t,\theta_0)^\intercal\right]\right)(\Lambda_T^{-1})^\intercal\\
&-\Lambda_T^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A^*_t(X_t,\theta_0)\varphi(Y_t,m(X_t,\theta_0))\varphi(Y_t,m(X_t,\theta_0))^\intercal A_t(X_t,\theta_0)^\intercal\right]\right)(\Delta^\intercal_{T,\mathbb{A}})^{-1}\\
&=\Delta^{-1}_{T,\mathbb{A}}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A_t(X_t,\theta_0)S_t(X_t,\theta_0)A_t(X_t,\theta_0)^\intercal\right]\right)(\Delta^\intercal_{T,\mathbb{A}})^{-1}
\end{aligned}$$

$$
+\Lambda_T^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[D_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}D_t(X_t,\theta_0)\right]\right)\Lambda_T^{-1}
$$

$$
-\Delta_{T,\mathbb{A}}^{-1}\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[A_t(X_t,\theta_0)D_t(X_t,\theta_0)\right]\right)\Lambda_T^{-1}-\Lambda_T\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[D_t(X_t,\theta_0)^\intercal A_t(X_t,\theta_0)^\intercal\right]\right)(\Delta_{T,\mathbb{A}}^\intercal)^{-1}
$$

$$
=\Delta_{T,\mathbb{A}}^{-1}\Sigma_{T,\mathbb{A}}\Delta_{T,\mathbb{A}}^{-1}-\Lambda_T^{-1}.
$$

As $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\chi_{t,T}\chi_{t,T}^\intercal\right]$ is positive semidefinite for all $T\geq 1$ as the sum of outer products, this shows that the matrix difference $\Delta_{T,\mathbb{A}}^{-1}\Sigma_{T,\mathbb{A}}\Delta_{T,\mathbb{A}}^{-1}-\Lambda_T^{-1}$ is positive semi-definite for all $T\geq 1$ and for all matrices $A_t(X_t,\theta_0)$.

For the proof of part (iii), assume that for some $t=1,\dots,T$, the matrix $A_t(X_t,\theta)$ is such that $A_t(X_t,\theta_0)\neq A^*_{t,C}(X_t,\theta_0)$ for any non-singular and deterministic matrix C with positive probability. Then, for some $t=1,\dots,T$, the matrix $M_{T,\mathbb{A}}(X_t,\theta_0):=\Delta_{T,\mathbb{A}}^{-1}A_t(X_t,\theta_0)-\Lambda_T^{-1}A^*_{t,C}(X_t,\theta_0)$ is nonzero with positive probability, as otherwise $A_t(X_t,\theta_0)=A^*_{t,\tilde{C}}(X_t,\theta_0)$ almost surely with $\tilde{C}=\Delta_{T,\mathbb{A}}\Lambda_T^{-1}C$. This implies that $M_{T,\mathbb{A}}(X_t,\theta_0)$ has positive rank with positive probability. Furthermore, the matrix $S_t(X_t,\theta_0)$ defined in (3.5) is positive definite with probability one for all $t=1,\dots,T$ by assumption. Consequently, we can apply the Cholesky decomposition and get that there exists a lower triangular matrix $G_t(X_t,\theta_0)$ with strictly positive diagonal entries such that $S_t(X_t,\theta_0)=G_t(X_t,\theta_0)G_t(X_t,\theta_0)^\intercal$ almost surely, i.e., the matrix $G_t(X_t,\theta_0)$ has full rank almost surely. Thus, the matrix $B_{T,\mathbb{A},t}(X_t,\theta_0):=M_{T,\mathbb{A}}(X_t,\theta_0)G_t(X_t,\theta_0)$ has positive rank for some $t=1,\dots,T$ with positive probability by Sylvester's rank inequality as it is the product of matrices with strictly positive rank (with positive probability) and full rank (almost surely). Consequently, there exists a $j\in\{1,\dots,k\}$ such that

$$
\mathbb{P}\big(B_{T,\mathbb{A},t}(X_t,\theta_0)^\intercal e_j\neq 0\big)>0,\qquad \text{for some } t=1,\dots,T, \tag{F.2}
$$

where e_j is the j-th standard basis vector of $\mathbb{R}^k$. Thus,

$$
e_j^\intercal\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\chi_{t,T}\chi_{t,T}^\intercal\right]\right)e_j=\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[e_j^\intercal M_{T,\mathbb{A}}(X_t,\theta_0)S_t(X_t,\theta_0)M_{T,\mathbb{A}}(X_t,\theta_0)^\intercal e_j\right]
$$

$$
=\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[e_j^\intercal M_{T,\mathbb{A}}(X_t,\theta_0)G_t(X_t,\theta_0)G_t(X_t,\theta_0)^\intercal M_{T,\mathbb{A}}(X_t,\theta_0)^\intercal e_j\right]
$$

$$
=\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[e_j^\intercal B_{T,\mathbb{A},t}(X_t,\theta_0)B_{T,\mathbb{A},t}(X_t,\theta_0)^\intercal e_j\right]=\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|B_{T,\mathbb{A},t}(X_t,\theta_0)^\intercal e_j\|^2\right]>0,
$$

for all $T\geq 1$, since all summands are non-negative and, invoking (F.2), at least one summand must be strictly positive, which shows that the matrix $\Delta_{T,\mathbb{A}}^{-1}\Sigma_{T,\mathbb{A}}\Delta_{T,\mathbb{A}}^{-1}-\Lambda_T^{-1}$ has at least one strictly positive eigenvalue, which concludes the proof of the theorem. □

Proof of Proposition 4.1. We first consider the case of the double moment functional. Straightforward calculations yield that the class of identification functions corresponding to the M-estimators based on loss functions given in (4.2) is given by

$$
\psi_{\phi_t}(Y_t,X_t,\theta)=A_{\phi_t}(X_t,\theta)\cdot\varphi_{\mathrm{mom}}(Y_t,m(X_t,\theta)), \tag{F.3}
$$

where φ is given in (4.1) and where $A_{\phi_t}(X_t,\theta)=\big(\nabla_\theta m_1(X_t,\theta)\quad \nabla_\theta m_2(X_t,\theta)\big)\cdot\nabla^2\phi_t\big(m(X_t,\theta)\big)$.

Applying Theorem 3.1 yields that the efficiency bound can be attained by a Z-estimator (and for equivalent M-estimators) if and only if $A^*_{t,C}(X_t,\theta) = CD_t(X_t,\theta)^\intercal S_t(X_t,\theta)^{-1}$ almost surely, where C is some deterministic and non-singular matrix, and where

$$S_t(X_t,\theta_0) = \mathbf{Var}_t\big((Y_t,Y_t^2)\big), \qquad \text{and} \qquad D_t(X_t,\theta_0) = \begin{pmatrix} \nabla_\theta m_1(X_t,\theta_0)^\intercal \\ \nabla_\theta m_2(X_t,\theta_0)^\intercal \end{pmatrix}.$$

By choosing $C = I_q$ and the strictly convex quadratic form $\phi_t(z) = \frac{1}{2} z^\intercal \mathbf{Var}_t\big((Y_t,Y_t^2)\big)^{-1} z$ for all $t \in \mathbb{N}$ and for all $z \in \mathbb{R}^2$, this yields that $\nabla^2 \phi_t\big(m(X_t,\theta_0)\big) = \mathbf{Var}_t\big((Y_t,Y_t^2)\big)^{-1}$ almost surely. Consequently, the M-estimator for the double moment regression is able to attain the efficient instrument matrix $A^*_{t,C}(X_t,\theta_0)$ (at θ_0) and consequently the Z-estimation efficiency bound.

For the situation of mean and variance, (F.3) takes the form $\psi_\phi(Y_t,X_t,\theta) = \tilde{A}_{t,\phi}(X_t,\theta) \cdot \varphi_{(\mathbb{E},\mathrm{Var})}(Y_t, m(X_t,\theta))$, where φ is given in (4.1) and where

$$\tilde{A}_{t,\phi}(X_t,\theta) = \begin{pmatrix} \nabla_\theta m_1(X_t,\theta)^\intercal \\ \nabla_\theta v(X_t,\theta)^\intercal + 2 m_1(X_t,\theta) \nabla_\theta m_1(X_t,\theta)^\intercal \end{pmatrix}^\intercal \cdot \nabla^2 \phi_t \begin{pmatrix} m_1(X_t,\theta) \\ v(X_t,\theta) + m_1^2(X_t,\theta) \end{pmatrix}.$$

Straight-forward calculations yield that $S_t(X_t,\theta_0) = \mathbf{Var}_t\big((Y_t,Y_t^2)\big)$ and

$$D_t(X_t,\theta_0) = \begin{pmatrix} \nabla_\theta m_1(X_t,\theta_0)^\intercal \\ \nabla_\theta v(X_t,\theta_0)^\intercal + 2 m_1(X_t,\theta_0) \nabla_\theta m_1(X_t,\theta_0)^\intercal \end{pmatrix}.$$

Thus, the efficient choice can be attained again with $\mathbb{R}^2 \ni z \mapsto \phi_t(z) = \frac{1}{2} z^\intercal \mathbf{Var}_t\big((Y_t,Y_t^2)\big)^{-1} z$. □

Proof of Theorem 4.2. Given that Assumptions (1) – (4) hold, Theorem 2.5 yields that any consistent M-estimator of semiparametric double quantile models is based on classical (strictly) consistent loss functions for the pair of two quantiles, given in (4.6). Furthermore, the M- and Z-estimator have identical asymptotic covariance if and only if the moment conditions of the Z- and derivative of the loss of the M-estimator coincide, or, respectively, their conditional expectations coincide, see discussion after (3.6). Thus, in the following we compare whether the derivatives of any strictly consistent loss function given in (4.6) can attain the efficient moment conditions of the Z-estimator almost surely.

We get that all identification functions which correspond to an M-estimator (in the form of a derivative of the conditional expectation almost surely) are given by

$$\psi_{g_{1,t},g_{2,t}}(Y_t,X_t,\theta) = \begin{pmatrix} \nabla_{\theta^\alpha} q_\alpha(X_t,\theta^\alpha) g'_{1,t}(q_\alpha(X_t,\theta^\alpha)) \big(\mathbb{1}_{\{Y_t \le q_\alpha(X_t,\theta^\alpha)\}} - \alpha\big) \\ \nabla_{\theta^\beta} q_\beta(X_t,\theta^\beta) g'_{2,t}(q_\beta(X_t,\theta^\beta)) \big(\mathbb{1}_{\{Y_t \le q_\beta(X_t,\theta^\beta)\}} - \beta\big) \end{pmatrix},$$

which can be written as $\psi_{g_{1,t},g_{2,t}}(Y_t,X_t,\theta) = A_{g_{1,t},g_{2,t}}(X_t,\theta) \varphi(Y_t, m(X_t,\theta))$, where

$$A_{g_{1,t},g_{2,t}}(X_t,\theta) = \begin{pmatrix} \nabla_{\theta^\alpha} q_\alpha(X_t,\theta^\alpha) g'_{1,t}(q_\alpha(X_t,\theta^\alpha)) & 0 \\ 0 & \nabla_{\theta^\beta} q_\beta(X_t,\theta^\beta) g'_{2,t}(q_\beta(X_t,\theta^\beta)) \end{pmatrix}. \tag{F.4}$$

We start by showing statement (B), assuming that the Z-estimation efficiency bound is attained by the M-estimator. From Theorem 3.1, part (i) and (ii), we get that the efficient instrument choice is given by $A^*_{t,C}(X_t,\theta_0) = CD_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$, at the true parameter θ_0, where C is some deterministic and nonsingular matrix and where $D_t(X_t,\theta_0)$ and $S_t(X_t,\theta_0)$ are given in (4.5). Furthermore, Theorem 3.1 part (iii) shows that any choice of $A_t(X_t,\theta_0)$ which deviates from

$A^*_{t,C}(X_t, \theta_0)$ (at the true parameter θ_0) with positive probability for some $t \in \mathbb{N}$, cannot attain the efficiency bound. Thus, in the following we show by contradiction that the general instrument matrix of the M-estimator, $A_{g_{1,t},g_{2,t}}(X_t, \theta_0)$, given in (F.4), cannot attain the necessary form $A^*_{t,C}(X_t, \theta_0)$ at the true parameter θ_0 with probability one for any deterministic matrix C.

For this, we assume that there exists a deterministic and non-singular $q \times q$ matrix C and functions $g_{1,t}$ and $g_{2,t}$ such that $A^*_{t,C}(X_t, \theta_0) = A_{g_{1,t},g_{2,t}}(X_t, \theta_0)$ almost surely for all $t \in \mathbb{N}$. We split $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ in its respective parts, where $C_{11} \in \mathbb{R}^{q_1 \times q_1}$, $C_{22} \in \mathbb{R}^{q_2 \times q_2}$, and $C_{12}, C_{21}^\intercal \in \mathbb{R}^{q_1 \times q_2}$. Then, the equation $A^*_{t,C}(X_t, \theta_0) = A_{g_{1,t},g_{2,t}}(X_t, \theta_0)$ is equivalent to

$$\begin{aligned} &\begin{pmatrix} \alpha(1-\alpha) g'_{1,t}(q_\alpha(X_t, \theta_0^\alpha)) \nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) & \alpha(1-\beta) g'_{1,t}(q_\alpha(X_t, \theta_0^\alpha)) \nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) \\ \alpha(1-\beta) g'_{2,t}(q_\beta(X_t, \theta_0^\beta)) \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta) & \beta(1-\beta) g'_{2,t}(q_\beta(X_t, \theta_0^\beta)) \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta) \end{pmatrix} \\ &= \begin{pmatrix} f_t(q_\alpha(X_t, \theta_0^\alpha)) C_{11} \nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) & f_t(q_\beta(X_t, \theta_0^\beta)) C_{12} \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta) \\ f_t(q_\alpha(X_t, \theta_0^\alpha)) C_{21} \nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) & f_t(q_\beta(X_t, \theta_0^\beta)) C_{22} \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta) \end{pmatrix}, \end{aligned} \tag{F.5}$$

which must hold element-wise for all four sub-components. Equality of the upper left component yields that there is some $A \in \mathcal{A}$ with $\mathbb{P}(A) = 1$ such that

$$\xi_t(\omega) \cdot \nabla_{\theta^\alpha} q_\alpha\big(X_t(\omega), \theta_0^\alpha\big) = C_{11} \cdot \nabla_{\theta^\alpha} q_\alpha\big(X_t(\omega), \theta_0^\alpha\big), \quad \forall \omega \in A \tag{F.6}$$

for the scalar random variable $\xi_t := \frac{\alpha(1-\alpha) g'_{1,t}(q_\alpha(X_t, \theta_0^\alpha))}{f_t(q_\alpha(X_t, \theta_0^\alpha))}$. Equation (F.6) is an Eigenvalue problem for the deterministic matrix C_{11} with stochastic Eigenvalues $\xi_t(\omega)$ and Eigenvectors $\nabla_{\theta^\alpha} q_\alpha(X_t(\omega), \theta_0^\alpha)$, $\omega \in A$. In the following, we show that this equation can hold only if ξ_t is constant on A.

By Assumption (DQ2), there are $\omega_1, \ldots, \omega_{q_1+1} \in A$ such that for $v_\ell := \nabla_{\theta^\alpha} q_\alpha(X_t(\omega_\ell), \theta_0^\alpha)$, $\ell \in \{1, \ldots, q_1+1\}$, any subset of cardinality q_1 of $\{v_1, \ldots, v_{q_1+1}\}$ is linearly independent. As C_{11} is a deterministic $q_1 \times q_1$ matrix, it can have at most q_1 different Eigenvalues. Let $\lambda_1, \ldots, \lambda_{q_1}$ be the Eigenvalues of C_{11} (not necessarily different, thus counted multiple times for higher algebraic multiplicities) ordered such that v_i is an Eigenvector for Eigenvalue λ_i for all $i = 1, \ldots, q_1$. Invoking that $v_1, \ldots, v_{q_1}$ are linearly independent, it holds that $\sum_{\lambda \in \{\lambda_1, \ldots, \lambda_{q_1}\}} \dim(E_\lambda) = q_1$, where the summation ignores repetitions in the set $\{\lambda_1, \ldots, \lambda_{q_1}\}$ and where E_λ denotes the Eigenspace corresponding to Eigenvalue λ. The Eigenvector v_{q_1+1} must be contained in E_{λ_i} for some $i = 1, \ldots, q_1$ as otherwise, the sum of the geometric multiplicities would exceed q_1. If $\dim(E_{\lambda_i}) = l < q_1$, then E_{λ_i} is spanned by l elements of $\{v_1, \ldots, v_{q_1}\}$, and as v_{q_1+1} is contained in E_{λ_i}, these l elements of $\{v_1, \ldots, v_{q_1}\}$ then must be linearly dependent together with v_{q_1+1}. This contradicts Assumption (DQ2). Thus, $\dim(E_{\lambda_i}) = q_1$ and consequently, the geometric multiplicity of λ_i is q_1, which then must equal the algebraic multiplicity. Hence, all Eigenvalues of C_{11} are equal, $\lambda_1 = \cdots = \lambda_{q_1}$, and consequently, ξ_t is constant on A, implying that it is constant almost surely. This implies that $g'_{1,t}(q_\alpha(X_t, \theta_0^\alpha)) = c_2 f_t(q_\alpha(X_t, \theta_0^\alpha))$ almost surely for some constant $c_2 > 0$ and for all $t \in \mathbb{N}$, i.e., (4.8). An analogous proof for the lower right entry of (F.5) shows (4.9), which concludes the proof of (B).

For (A) we start with the 'only if' direction assuming that the M-estimator attains the efficiency bound. From part (B), we already obtain that (4.8) and (4.9) must hold. Exploiting $\nabla_{\theta^\alpha} q_\alpha(X_t, \theta_0^\alpha) = \nabla_{\theta^\beta} q_\beta(X_t, \theta_0^\beta)$ and $g'_{1,t}(q_\alpha(X_t, \theta_0^\alpha)) = c_2 f_t(q_\alpha(X_t, \theta_0^\alpha))$, the upper right component of (F.5) implies that

$$\frac{\alpha(1-\beta)c_2 f_t(q_\alpha(X_t,\theta_0^\alpha))}{f_t(q_\beta(X_t,\theta_0^\beta))}\cdot\nabla_{\theta^\alpha}q_\alpha(X_t,\theta_0^\alpha)=C_{12}\cdot\nabla_{\theta^\alpha}q_\alpha(X_t,\theta_0^\alpha), \tag{F.7}$$

almost surely. Applying the same Eigenvalue argument to (F.7) (recalling that $\nabla_{\theta^\alpha}q_\alpha(X_t,\theta_0^\alpha)=\nabla_{\theta^\beta}q_\beta(X_t,\theta_0^\beta)$ implies that $q_1=q_2$ such that C_{12} is quadratic) yields (4.7).

For the 'if' implication in (A), we assume that (4.7), (4.8) and (4.9) hold. We choose $C_{11}=\alpha(1-\alpha)c_2 I_{q_1\times q_1}$, $C_{12}=\alpha(1-\beta)c_1c_2 I_{q_1\times q_2}$, $C_{21}=\alpha(1-\beta)c_3/c_1 I_{q_2\times q_1}$ and $C_{22}=\beta(1-\beta)c_3 I_{q_2\times q_2}$, where $\det(C)\neq 0$ follows from $0<\alpha<\beta<1$. Thus, straightforward calculations yield that $A_{g_{1,t},g_{2,t}}(X_t,\theta_0)=A^*_{t,C}(X_t,\theta_0)$ holds almost surely for all $t\in\mathbb{N}$. Applying Theorem 3.1 yields the claim. □

Proof of Theorem 4.3. This proof follows the general ideas of the proof of Theorem 4.2. Given Assumptions (1)–(4), Theorem 2.5 yields that any consistent M-estimator of semiparametric joint quantile and ES models is based on classical (strictly) consistent loss functions given in (4.18). Thus, in the following we compare whether the derivatives of any strictly consistent loss function are able to attain the efficient moment conditions of Theorem 3.1 almost surely.

We get that all identification functions which correspond to an M-estimator (in the form of a derivative almost surely) are given by $\psi_{g_t,\phi_t}\big(Y_t,X_t,\theta\big)$ equalling

$$\begin{pmatrix}\nabla_{\theta^q}q_\alpha(X_t,\theta^q)\big(g_t'(q_\alpha(X_t,\theta^q))+\phi_t'(e_\alpha(X_t,\theta^e))/\alpha\big)\left(\mathbb{1}_{\{Y_t\le q_\alpha(X_t,\theta^q)\}}-\alpha\right)\\ \nabla_{\theta^e}e_\alpha(X_t,\theta^e)\phi_t''(e_\alpha(X_t,\theta^e))\left(e_\alpha(X_t,\theta^e)-q_\alpha(X_t,\theta^q)+\frac{1}{\alpha}(q_\alpha(X_t,\theta^q)-Y_t)\mathbb{1}_{\{Y_t\le q_\alpha(X_t,\theta^q)\}}\right)\end{pmatrix}.$$

This implies that the moment conditions corresponding to an M-estimator can be written as $\psi_{g_t,\phi_t}\big(Y_t,X_t,\theta\big)=A_{g_t,\phi_t}(X_t,\theta)\varphi\big(Y_t,m(X_t,\theta)\big)$, where $\varphi\big(Y_t,m(X_t,\theta)\big)$ is given in (4.15), and

$$A_{g_t,\phi_t}(X_t,\theta)=\begin{pmatrix}\big(g_t'(q_\alpha(X_t,\theta^q))+\phi_t'(e_\alpha(X_t,\theta^e))/\alpha\big)\nabla_{\theta^q}q_\alpha(X_t,\theta^q) & 0\\ 0 & \phi_t''(e_\alpha(X_t,\theta^e))\nabla_{\theta^e}e_\alpha(X_t,\theta^e)\end{pmatrix}.$$

To show (B), we assume that the Z-estimation efficiency bound is attained by the M-estimator. From Theorem 3.1, we get that the efficient estimator has to fulfill the condition $A^*_{t,C}(X_t,\theta_0)=CD_t(X_t,\theta_0)^\intercal S_t(X_t,\theta_0)^{-1}$ for some deterministic and nonsingular matrix C, where $D_t(X_t,\theta_0)$ and $S_t(X_t,\theta_0)$ are given in (4.16) and (4.17). Thus, we verify whether there exists a deterministic and non-singular $q\times q$ matrix C (and appropriate functions g_t and ϕ_t) such that $A^*_{t,C}(X_t,\theta_0)=A_{g_t,\phi_t}(X_t,\theta_0)$ almost surely, i.e., whether $CD_t(X_t,\theta_0)^\intercal=A_{g_t,\phi_t}(X_t,\theta_0)S_t(X_t,\theta_0)$ holds almost surely. By splitting the matrix $C=\begin{pmatrix}C_{11} & C_{12}\\ C_{21} & C_{22}\end{pmatrix}$ in its respective parts, where $C_{11}\in\mathbb{R}^{q_1\times q_1}$, $C_{22}\in\mathbb{R}^{q_2\times q_2}$, and $C_{12},C_{21}^\intercal\in\mathbb{R}^{q_1\times q_2}$, this simplifies to the following four equalities,

$$C_{11}\nabla_{\theta^q}q_\alpha(X_t,\theta_0^q)=(1-\alpha)\frac{\alpha g_t'(q_\alpha(X_t,\theta_0^q))+\phi_t'(e_\alpha(X_t,\theta_0^e))}{f_t(q_\alpha(X_t,\theta_0^q))}\nabla_{\theta^q}q_\alpha(X_t,\theta_0^q), \tag{F.8}$$

$$\begin{aligned}C_{12}\nabla_{\theta^e}e_\alpha(X_t,\theta_0^e)=(1-\alpha)&\big(q_\alpha(X_t,\theta_0^q)-e_\alpha(X_t,\theta_0^e)\big)\\ &\times\left(g_t'(q_\alpha(X_t,\theta_0^q))+\phi_t'(e_\alpha(X_t,\theta_0^e))/\alpha\right)\nabla_{\theta^q}q_\alpha(X_t,\theta_0^q),\end{aligned} \tag{F.9}$$

$$C_{21}\nabla_{\theta^q}q_\alpha(X_t,\theta_0^q)=\frac{(1-\alpha)\big(q_\alpha(X_t,\theta_0^q)-e_\alpha(X_t,\theta_0^e)\big)\phi_t''(e_\alpha(X_t,\theta_0^e))}{f_t(q_\alpha(X,\theta_0^q))}\nabla_{\theta^e}e_\alpha(X_t,\theta_0^e), \tag{F.10}$$

$$
\begin{aligned}
&C_{22}\nabla_{\theta^e} e_\alpha(X_t,\theta_0^e) = \phi_t''(e_\alpha(X_t,\theta_0^e))\nabla_{\theta^e} e_\alpha(X_t,\theta_0^e) \\
&\quad\times\left(\frac{1}{\alpha}\operatorname{Var}_t\left(Y_t \middle| Y_t \le q_\alpha(X_t,\theta_0^q)\right) + \frac{1-\alpha}{\alpha}\left(e_\alpha(X_t,\theta_0^e) - q_\alpha(X_t,\theta_0^q)\right)^2\right),
\end{aligned} \tag{F.11}
$$

which have to hold almost surely. Using the same Eigenvalue argument as in the proof of Theorem 4.2, equation (F.8) implies that

$$
(1-\alpha)\big(\alpha g_t'(q_\alpha(X_t,\theta_0^q)) + \phi_t'(e_\alpha(X_t,\theta_0^e))\big) = \tilde{c}_1 f_t(q_\alpha(X_t,\theta_0^q)) \tag{F.12}
$$

almost surely for some constant $\tilde{c}_1 > 0$. Equation (4.24) follows by setting $c_6 = \tilde{c}_1/(\alpha(1-\alpha))$. Similarly, (F.11) implies that

$$
\frac{\tilde{c}_2}{\phi_t''(e_\alpha(X_t,\theta_0^e))} = \frac{1}{\alpha}\operatorname{Var}_t\left(Y_t \middle| Y_t \le q_\alpha(X_t,\theta_0^q)\right) + \frac{1-\alpha}{\alpha}\left(e_\alpha(X_t,\theta_0^e) - q_\alpha(X_t,\theta_0^q)\right)^2 \tag{F.13}
$$

almost surely for some constant $\tilde{c}_2 > 0$. Furthermore, combining (F.9) and (F.10) implies

$$
\begin{aligned}
&C_{12}C_{21}\nabla q_\alpha(X_t,\theta_0^q) = \nabla q_\alpha(X_t,\theta_0^q)\frac{(1-\alpha)^2}{\alpha} \\
&\quad\times\frac{\left(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\right)^2\phi_t''(e_\alpha(X_t,\theta_0^e))\left(\alpha g_t'(q_\alpha(X_t,\theta_0^q)) + \phi_t'(e_\alpha(X_t,\theta_0^e))\right)}{f_t(q_\alpha(X_t,\theta_0^q))}
\end{aligned}
$$

almost surely and employing the same Eigenvalue argument again yields that

$$
\begin{aligned}
&\left(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\right)^2\phi_t''(e_\alpha(X_t,\theta_0^e))\left(\alpha g_t'(q_\alpha(X_t,\theta_0^q)) + \phi_t'(e_\alpha(X_t,\theta_0^e))\right) \\
&= \frac{\tilde{c}_3\alpha}{(1-\alpha)^2} f_t(q_\alpha(X_t,\theta_0^q))
\end{aligned} \tag{F.14}
$$

almost surely for some constant $\tilde{c}_3 > 0$. Substituting (F.12) and (F.13) into (F.14) finally yields that

$$
\operatorname{Var}_t\left(Y_t \middle| Y_t \le q_\alpha(X_t,\theta_0^q)\right) = (1-\alpha)\left(\frac{\tilde{c}_1\tilde{c}_2}{\tilde{c}_3} - 1\right)\left(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\right)^2. \tag{F.15}
$$

almost surely. By defining the constant $c_1 := (1-\alpha)\left(\frac{\tilde{c}_1\tilde{c}_2}{\tilde{c}_3} - 1\right)$, we obtain (4.19), where the positivity of c_1 follows from that fact that both sides of (F.15) are positive. Substituting (F.15) into (F.13) yields (4.21), which concludes the proof of statement (B).

For (A) we start with the 'only if' direction, assuming that the M-estimator attains the efficiency bound. From part (B) we obtain that (4.19), (4.21) and (4.24) must hold. Employing the same Eigenvalue argument as before, we obtain from (F.10) that $\tilde{c}_4 f_t(q_\alpha(X_t,\theta_0^q))/(1-\alpha) = \big(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\big)\phi_t''(e_\alpha(X_t,\theta_0^e))$ for some constant $\tilde{c}_4 > 0$, where we additionally exploited that $\nabla_{\theta^q} q_\alpha(X_t,\theta_0^q) = \nabla_{\theta^e} e_\alpha(X_t,\theta_0^e)$ almost surely. Combining (4.19) and (4.21) yields $\phi_t''(e_\alpha(X_t,\theta_0^e)) = c_3/c_1\big(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\big)^{-2}$, which in turn leads us to

$$
f_t(q_\alpha(X_t,\theta_0^q)) = \frac{c_2}{q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)}, \tag{F.16}
$$

almost surely for where $c_2 = (1-\alpha)c_3/(c_1\tilde{c}_4) > 0$, establishing (4.20). Using again that $\phi_t''(e_\alpha(X_t,\theta_0^e)) = c_3/c_1\big(q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)\big)^{-2}$ and since the support of $e_\alpha(X_t,\theta_0^e)$ is a non-degenerate interval by assumption, it must hold that

$$\phi_t'(e_\alpha(X_t,\theta_0^e)) = \frac{c_3/c_1}{q_\alpha(X_t,\theta_0^q) - e_\alpha(X_t,\theta_0^e)} + \tilde{c}_{5,t}, \tag{F.17}$$

almost surely for some deterministic, but possibly time-varying constant $\tilde{c}_{5,t} \in \mathbb{R}$ for all $t \in \mathbb{N}$. Combining (F.12), (F.16) and (F.17) yields that

$$g_t'(q_\alpha(X_t,\theta_0^q)) = c_4 f_t(q_\alpha(X_t,\theta_0^q)) + c_{5,t}, \tag{F.18}$$

where $c_4 := \left(\frac{\tilde{c}_1}{\alpha(1-\alpha)} - \frac{c_3}{\alpha c_1 c_2}\right) \in \mathbb{R}$ and $c_{5,t} := -\tilde{c}_{5,t}/\alpha$, which establishes (4.22). Eventually, employing (F.12) and (F.18) yields that $\phi_t'(e_\alpha(X_t,\theta_0^e))/\alpha = \tilde{c}_1 f_t(q_\alpha(X_t,\theta_0^q))/(\alpha(1-\alpha)) - c_4 f_t(q_\alpha(X_t,\theta_0^q)) - c_{5,t}$, and hence $\phi_t'(e_\alpha(X_t,\theta_0^e)) = c_3 f_t(q_\alpha(X_t,\theta_0^q))/(c_1c_2) - \alpha c_{5,t}$, which shows (4.23) and concludes this direction.

For the 'if' implication in statement (A), we assume the conditions (4.19)–(4.23). Choosing $C_{11} = \alpha(1-\alpha)\left(c_4 + \frac{c_3}{\alpha c_1 c_2}\right) I_{q_1\times q_1}$, $C_{12} = \frac{(1-\alpha)}{\alpha c_1 c_3^2}(\alpha c_1 c_2 c_4 + c_3) I_{q_1 \times q_2}$, $C_{21} = \frac{(1-\alpha)c_3}{c_1 c_2} I_{q_2\times q_1}$ and $C_{22} = \frac{c_1+1-\alpha}{\alpha c_1 c_3} I_{q_2\times q_2}$, automatically yields $\det(C) \neq 0$, and straight-forward calculations yield that (F.8)–(F.11) are satisfied and thus, $A_{g_t,\phi_t}(X_t,\theta_0) = A^*_{t,C}(X_t,\theta_0)$ holds almost surely for all $t \in \mathbb{N}$. Applying Theorem 3.1 yields the claim □

Proof of Theorem B.1. The assertion at (i) is a direct consequence of the linearity of the expectation. For (ii), the proof of the existence of h follows along the lines of Theorem 3.2 in Fissler and Ziegel (2016). One just replaces $\nabla \bar{S}(x,F)$ with $\bar{\varphi}'(F,\xi)$ and $\bar{V}(x,F)$ with $\varphi(F,\xi)$. Note also that the surjectivity of Γ onto $\operatorname{int}(\Xi)$ is implicitly given by Assumption (5) and the fact that φ is a strict $\mathcal{F}$-identification function for Γ. If φ' satisfies Assumption (5) as well, one directly obtains that h must have full rank on $\operatorname{int}(\Xi)$. If the expected identification functions are both continuous, the continuity of h follows again exactly like in the proof of Theorem 3.2 in Fissler and Ziegel (2016). Finally, the pointwise assertion at (B.1) follows like in the proof of Proposition 3.4 *ibidem*; see the supplemental material to Fissler and Ziegel (2016) and Fissler and Ziegel (2019a). □

Proof of Proposition B.4. Clearly, the tower property implies $\mathbb{E}\big[A(X,\theta_0)\varphi\big(Y,m(X,\theta_0)\big)\big] = 0$. For $\theta \neq \theta_0$ the mean value theorem yields

$$\begin{aligned}\mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)|X\big] &= \mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)|X\big] - \mathbb{E}\big[\varphi\big(Y,m(X,\theta_0)\big)|X\big] \\ &= \nabla_\theta \mathbb{E}\big[\varphi\big(Y,m(X,\theta)\big)|X\big]\big|_{\theta=\theta'}(\theta-\theta_0) = D(X,\theta')(\theta-\theta_0)\end{aligned}$$

Therefore $\mathbb{E}\big[A(X,\theta)\varphi\big(Y,m(X,\theta)\big)\big] = \mathbb{E}\big[A(X,\theta)D(X,\theta')\big](\theta-\theta_0) \neq 0$. □

Proof of Proposition B.5. Assume that ψ is not a strict conditional $\mathcal{F}_{\mathcal{Z}}$-model-identification function for θ_0. That means there is $Z = (Y,X) \in \mathcal{Z}$ with true model parameter $\theta_0 = \theta_0(F_Z)$ such that

$$\mathbb{P}\big(\mathbb{E}[\psi(Y,X,\theta_0)|X] \neq 0\big) > 0 \tag{F.19}$$

or

$$\exists \theta' \neq \theta_0 \colon \mathbb{E}[\psi(Y,X,\theta')|X] = 0 \quad \text{a.s.} \tag{F.20}$$

If (F.20) holds, then we can directly apply the tower property to obtain that for some $\theta' \neq \theta_0$ we have $\mathbb{E}[\psi_A(Y,X,\theta')] = \mathbb{E}\big[A(X,\theta')\mathbb{E}[\psi(Y,X,\theta')|X]\big] = 0$, which means that ψ_A is not a strict unconditional $\mathcal{F}_{\mathcal{Z}}$-identification function for θ_0. Now, we assume that (F.19) holds. Then using (B.3) we can conclude that $\mathbb{P}\big(\mathbb{E}[\psi_A(Y,X,\theta_0)|X] \neq 0\big) = \mathbb{P}\big(A(X,\theta_0)\mathbb{E}[\psi(Y,X,\theta_0)|X] \neq 0\big) \geq \mathbb{P}\big(\{\text{rank}(A(X,\theta_0)) = k\} \cap \{\mathbb{E}[\psi(Y,X,\theta_0)|X] \neq 0\}\big) > 0$. Then, we can again argue that there exists a component $j \in \{1,\dots,q\}$ such that $\mathbb{P}\big(\mathbb{E}[\psi_{A,j}(Y,X,\theta_0)|X] < 0\big) > 0$ or $\mathbb{P}\big(\mathbb{E}[\psi_{A,j}(Y,X,\theta_0)|X] > 0\big) > 0$, and we continue as in the proof of Theorem 2.5 (iv). □

Proof of Proposition C.1. It holds that $\mathbb{E}\left[A_t^*(X_t,\theta_0)\varphi\big(Y_t,m(X_t,\theta_0)\big)\right] = 0$ since we have that $\mathbb{E}\left[\varphi\big(Y_t,m(X_t,\theta_0)\big)\big|X_t\right] = 0$. The reverse direction is a little more involved. For this, straightforward calculations yield that for any $\theta \in \Theta$

$$\mathbb{E}\left[A_t^*(X_t,\theta_0)\varphi\big(Y_t,m(X_t,\theta)\big)\right] = \mathbb{E}\left[U_1\nabla_\theta q_\alpha(X_t,\theta_0^\alpha) + U_2\nabla_\theta q_\beta(X_t,\theta_0^\beta)\right],$$

where the scalar and $\sigma(X_t)$-measurable random variables U_1 and U_2 are given by

$$U_1 = \frac{f_t(q_\alpha(X_t,\theta_0^\alpha))}{\alpha(1-\alpha)\beta - \alpha^2(1-\beta)}\big(\beta a - \alpha b\big) \qquad \text{and}$$
$$U_2 = \frac{f_t(q_\beta(X_t,\theta_0^\beta))}{\beta(1-\alpha)(1-\beta) - \alpha(1-\beta)^2}\big(-(1-\beta)a + (1-\alpha)b\big),$$

with $a = F_t(q_\alpha(X_t,\theta^\alpha)) - \alpha$ and $b = F_t(q_\beta(X_t,\theta^\beta)) - \beta$. As $\nabla_\theta q_\alpha(X_t,\theta_0^\alpha) = \begin{pmatrix} X_t \\ 0 \end{pmatrix}$ and $\nabla_\theta q_\beta(X_t,\theta_0^\beta) = \begin{pmatrix} 0 \\ X_t \end{pmatrix}$, it holds that $\mathbb{E}\left[A_t^*(X_t,\theta_0)\varphi\big(Y_t,m(X_t,\theta)\big)\right] = 0$ if and only if

$$\mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))\big(\beta a - \alpha b\big)X_t\right] = 0 \quad \text{and} \quad \mathbb{E}\left[f_t(q_\beta(X_t,\theta_0^\beta))\big((1-\beta)a - (1-\alpha)b\big)X_t\right] = 0. \tag{F.21}$$

As $f_t(q_\alpha(X_t,\theta_0^\alpha)) = c_t f_t(q_\beta(X_t,\theta_0^\beta))$ almost surely by assumption (where c_t is deterministic), this implies that

$$\beta\mathbb{E}\big[f_t(q_\alpha(X_t,\theta_0^\alpha))aX_t\big] - \alpha\mathbb{E}\big[f_t(q_\alpha(X_t,\theta_0^\alpha))bX_t\big] = 0 \qquad \text{and}$$
$$c_t(1-\beta)\mathbb{E}\big[f_t(q_\alpha(X_t,\theta_0^\alpha))aX_t\big] - c_t(1-\alpha)\mathbb{E}\big[f_t(q_\alpha(X_t,\theta_0^\alpha))bX_t\big] = 0.$$

Solving this system of equations, where we exploit that $c_t \neq 0$, and combining it with (F.21) and the fact that $\alpha \neq \beta$, we arrive at

$$\mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))aX_t\right] = 0 \quad \text{and} \quad \mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))bX_t\right] = 0. \tag{F.22}$$

We now proceed by a proof through contradiction with an argument similar as in Dimitriadis and Bayer (2019). For this, assume that $\theta \neq \theta_0$. Using the zero-condition in (F.22), we get

$$\begin{aligned}
0 &= \mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))aX_t^\intercal\right]\left(\theta^\alpha - \theta_0^\alpha\right) \\
&= \mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))X_t^\intercal\left(\theta^\alpha - \theta_0^\alpha\right)\left(F_t(q_\alpha(X_t,\theta^\alpha)) - F_t(q_\alpha(X_t,\theta_0^\alpha))\right)\right] \\
&= \mathbb{E}\left[f_t(q_\alpha(X_t,\theta_0^\alpha))f_t(q_\alpha(X_t,\tilde{\theta}^\alpha))\left(X_t^\intercal\left(\theta^\alpha - \theta_0^\alpha\right)\right)^2\right],
\end{aligned}$$

where we have used the mean value theorem and the linearity of the model to obtain the last identity and where $\tilde{\theta}^\alpha = (1-\lambda)\theta_0^\alpha + \lambda\theta^\alpha$ for some $\lambda \in [0,1]$. By assumption, the density is strictly positive such that we can conclude that $\mathbb{P}(X_t^\intercal(\theta^\alpha - \theta_0^\alpha) = 0) = 1$. Then, due to Assumption (1), it must hold that $\theta^\alpha = \theta_0^\alpha$. Employing a similar argument to θ^β yields that $\theta^\beta = \theta_0^\beta$, which concludes this proof. □

References

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic Efficiency of Semiparametric Two-step GMM. *Rev. Econ. Stud.*, 81(3):919–943.

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable growth. *Am. Econ. Rev.*, 109(4):1263–89.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Andrews, D. (1994). Empirical Process Methods in Econometrics. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 37, pages 2247–2294. Elsevier.

Angrist, J., Chernozhukov, V., and Fernandez-Val, I. (2006). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica*, 74(2):539–563.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.*, 12(2):171–178.

Barendse, S. (2020). Efficiently Weighted Estimation of Tail and Interquartile Expectations. *Preprint.* Available at https://dx.doi.org/10.2139/ssrn.2937665.

Bartalotti, O. (2013). GMM Efficiency and IPW Estimation for Nonsmooth Functions. *Preprint.* Available at http://repec.tulane.edu/RePEc/pdf/tul1301.pdf.

Basel Committee (2016). Minimum capital requirements for Market Risk. Technical report, Bank for International Settlements. Available at http://www.bis.org/bcbs/publ/d352.pdf.

Basel Committee (2019). Minimum capital requirements for Market Risk. Technical report, Bank for International Settlements. Available at https://www.bis.org/bcbs/publ/d457.pdf.

Bayer, S. and Dimitriadis, T. (2020). Regression based expected shortfall backtesting. *Journal of Financial Econometrics (forthcoming).* Available at https://doi.org/10.1093/jjfinec/nbaa013.

Bellini, F. and Bignozzi, V. (2015). On elicitable risk measures. *Quant. Finance*, 15(5):725–733.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins series in the mathematical sciences. Springer New York.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica*, 58:1443–1458.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econom.*, 31(3):307–327.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2020). Evaluating epidemic forecasts in an interval format. *Preprint.* Available at https://arxiv.org/abs/2005.12881.

Brehmer, J. (2017). Elicitability and its application in risk management. Master's thesis, University of Mannheim. Available at https://arxiv.org/abs/1707.09604.

Brehmer, J. and Gneiting, T. (2020). Scoring Interval Forecasts: Equal-Tailed, Shortest, and Modal Interval. *Preprint.* Available at https://arxiv.org/abs/2007.05709.

Brown, B. W. (1983). The identification problem in systems nonlinear in the variables. *Econometrica*, 51(1):175–196.

Buchinsky, M. (1994). Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression. *Econometrica*, 62(2):405–458.

Catania, L. and Luati, A. (2019). Semiparametric modeling of multiple quantiles. *Preprint.* Available at http://dx.doi.org/10.2139/ssrn.3494995.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econom.*, 34(3):305–334.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Couperier, O. and Leymarie, J. (2019). Backtesting expected shortfall via multi-quantile regression. *Preprint.* Available at https://halshs.archives-ouvertes.fr/halshs-01909375v5.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *J. Appl. Econ.*, 28(5):777–795.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians.* Advanced Texts in Econometrics. OUP Oxford.

Dimitriadis, T. and Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electron. J. Statist.*, 13(1):1823–1871.

Dimitriadis, T., Liu, X., and Schnaitmann, J. (2020). Encompassing Tests for Value at Risk and Expected Shortfall Multi-Step Forecasts based on Inference on the Boundary. *Preprint.* Available at https://arxiv.org/abs/2009.07341.

Dimitriadis, T., Patton, A. J., and Schmidt, P. (2019). Testing Forecast Rationality for Measures of Central Tendency. *Preprint.* Available at https://arxiv.org/abs/1910.12545.

Dimitriadis, T. and Schnaitmann, J. (2020). Forecast Encompassing Tests for the Expected Shortfall. *Int. J. Forecast. (forthcoming).* Available at https://doi.org/10.1016/j.ijforecast.2020.07.008.

Domínguez, M. A. and Lobato, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, 72(5):1601–1615.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statist. Sinica*, 1(1):93–125.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *J. Bus. Econom. Statist.*, 22(4):367–381.

Fissler, T. (2017). *On Higher Order Elicitability and Some Limit Theorems on the Poisson and Wiener Space.* PhD thesis, University of Bern. Available at http://biblio.unibe.ch/download/eldiss/17fissler_t.pdf.

Fissler, T., Frongillo, R., Hlavinová, J., and Rudloff, B. (2020). Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Preprint.* Available at https://arxiv.org/abs/1910.07912v2.

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.*, 44(4):1680–1707.

Fissler, T. and Ziegel, J. F. (2019a). Erratum: Higher Order Elicitability and Osband's Principle. *Preprint.* Available at https://arxiv.org/abs/1901.08826v1.

Fissler, T. and Ziegel, J. F. (2019b). Order-sensitivity and equivariance of scoring functions. *Electron. J. Stat.*, 13(1):1166–1211.

Gerlach, R. and Wang, C. (2020a). Bayesian Semi-parametric Realized-CARE Models for Tail Risk Forecasting Incorporating Realized Measures. *Preprint.* Available at https://arxiv.org/abs/1612.08488.

Gerlach, R. and Wang, C. (2020b). Semi-parametric dynamic asymmetric laplace models for tail risk forecasting, incorporating realized measures. *Int. J. Forecast.*, 36(2):489 – 506.

Giacomini, R. and Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts. *J. Bus. Econom. Statist.*, 23:416–431.

Gneiting, T. (2011a). Making and Evaluating Point Forecasts. *J. Amer. Statist. Assoc.*, 106:746–762.

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *Int. J. Forecast.*, 27(2):197–207.

Gneiting, T. and Raftery, A. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.*, 102:359–378.

Gourieroux, C. and Jasiak, J. (2008). Dynamic quantile models. *J. Econom.*, 147(1):198–205.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700.

Hall, A. (2005). *Generalized Method of Moments.* Advanced texts in econometrics. Oxford University Press.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–54.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *J. Econom.*, 30(1):203 – 238.

Hristache, M. and Patilea, V. (2016). Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. *Econom. Theory*, 32(4):917–946.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. Berkeley: University of California Press.

Janková, J. and van de Geer, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *Ann. Statist.*, 46(5):2336–2359.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, pages 33–50.

Koenker, R. and Xiao, Z. (2006). Quantile Autoregression. *J. Amer. Statist. Assoc.*, 101(475):980–990.

Komunjer, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *J. Econom.*, 128(1):137–164.

Komunjer, I. (2012). Global identification in nonlinear models with moment restrictions. *Econom. Theory*, 28(4):719–729.

Komunjer, I. and Vuong, Q. (2010a). Efficient estimation in dynamic conditional quantile models. *J. Econom.*, 157(2):272–285.

Komunjer, I. and Vuong, Q. (2010b). Semiparametric efficiency bound in time-series models for conditional quantiles. *Econom. Theory*, 26(02):383–405.

Königsberger, K. (2004). *Analysis 2*. Springer-Verlag, Berlin Heidelberg New York, 5th edition.

Lambert, N. (2013). Elicitation and Evaluation of Statistical Functionals. *Preprint*. Available at http://web.stanford.edu/~nlambert/papers/elicitation.pdf.

Lambert, N., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, Chicago, Il, USA. ACM.

Meng, X. and Taylor, J. W. (2020). Estimating value-at-risk and expected shortfall using the intraday low and range data. *Eur. J. Oper. Res.*, 280(1):191 – 202.

Murphy, A. H. and Daan, H. (1985). Forecast Evaluation. In Murphy, A. H. and Katz, R. W., editors, *Probability, Statistics and Decision Making in the Atmospheric Sciences*, pages 379–437. Westview Press, Boulder, Colorado.

Nau, R. F. (1985). Should Scoring Rules Be 'Effective'? *Manag. Sci.*, 31(5):527–535.

Newey, W. K. (1985). Maximum Likelihood Specification Testing and Conditional Moment Tests. *Econometrica*, 53(5):1047–1070.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econ.*, 5(2):99–135.

Newey, W. K. (1993). Efficient Estimation of Models with Conditional Moment Restrictions. In Maddala, G., Rao, C., and Vinod, H., editors, *Handbook of Statistics, Volume 11: Econometrics*.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Newey, W. K. and Smith, R. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.*, 11(4):1833–1874.

Osband, K. H. (1985). *Providing Incentives for Better Cost Forecasting*. PhD thesis, University of California, Berkeley.

Patton, A. (2011). Volatility forecast comparison using imperfect volatility proxies. *J. Econom.*, 160(1):246–256.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *J. Econom.*, 211(2):388 – 413.

Petropoulos, F. and Makridakis, S. (2020). Forecasting the novel coronavirus covid-19. *PLOS ONE*, 15(3):1–8.

Prokhorov, A. and Schmidt, P. (2009). GMM redundancy results for general missing data problems. *J. Econom.*, 151(1):47 – 55.

Roehrig, C. S. (1988). Conditions for identification in nonparametric and parametric models. *Econometrica*, 56(2):433–447.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3):577–591.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *J. Amer. Statist. Assoc.*, 66:783–801.

Schmidt, L. and Zhu, Y. (2016). Quantile spacings: A simple method for the joint estimation of multiple quantiles without crossing. *Preprint*. Available at http://dx.doi.org/10.2139/ssrn.2220901.

Spady, R. and Stouli, S. (2018). Simultaneous mean-variance regression. *Preprint*. Available at https://arxiv.org/abs/1804.01631.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 187–195, Berkeley, Calif. University of California Press.

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. (2014). Elicitation and Identification of Properties. *JMLR Workshop Conf. Proc.*, 35:1–45.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *J. Bus. Econom. Statist.*, 37(1):121–133.

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *Int. J. Forecast.*, 36(2):428–441.

UMass-Amherst Influenza Forecasting Center of Excellence (2020). COVID-19 Forecast Hub. Available at https://github.com/reichlab/covid19-forecast-hub.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, R. and Wei, Y. (2020). Risk functionals with convex level sets. *Math. Finance*, 30(4):1337–1367.

Weber, S. (2006). Distribution-Invariant Risk Measures, Information, and Dynamic Consistency. *Math. Finance*, 16:419–441.

Weiss, A. A. (1991). Estimating nonlinear dynamic models using least absolute error estimation. *Econom. Theory*, 7(01):46–68.

White, H. (2001). *Asymptotic Theory for Econometricians*. Academic Press, San Diego.

White, H., Kim, T.-H., and Manganelli, S. (2015). VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *J. Econom.*, 187(1):169–188.