Estimating Long Tail Models for Risk Trends

Saahil Shenoy
Department of Physics
Stanford University, Stanford, CA
saahils@stanford.edu

Dimitry Gorinevsky
Department of Electrical Engineering
Stanford University, Stanford, CA
gorin@stanford.edu

***Abstract*—This paper develops a method for estimating trends of extreme events statistics across multiple time periods. Some of the periods might have no extreme events and some might have much data. The extreme event distribution is modeled with a Pareto or exponential tail. The method requires selecting an extreme event threshold and then solving two convex problems for the tail parameters. Solving one provides a smoothed tail rate trend, solving another, the smoothed trend of the tail quantile level. The approach is illustrated by trending the 10-year extreme event risks for S&P 500 index daily losses and for peak power load in electrical utility data.**

I. INTRODUCTION

Long tail distributions appear in many signal processing problems such as internet traffic analysis [1], image processing [2], [3], and risk in finance [4]. Detection and decision algorithms for these problems require distribution knowledge.

The Peaks Over Threshold (POT) exceedance data is usually sparse and simple parametric fit procedures, such as the Hill's estimator for the Pareto tail model, are used [5], [6], [7], [8]. To improve estimation, [9], [10] propose using a conjugate Bayesian prior, the Gamma distribution. The prior distributions that are long tailed can also reduce sensitivities to spurious outliers, see [11], [12].

In modern signal processing problems, there could be thousands of one-per million extreme events. This affords more detailed extreme statistics modeling, such as estimating trends of tail parameters, the subject of this paper. The trends of extreme event risks are of much interest in problems of climate science, finance, industrial data, and actuarial risk.

There seems to be little prior work on multi-period estimation for the tail models. One ad hoc approach assumes that tail distribution parameters depend on time in accordance with a given regression model. In [13], this is a linear trend model. In [14], a more complex regression is used. The non-convex problems in [13], [14] are computationally difficult. The multi-period model in [15] uses a Bayesian prior for the exceedance number, but not for the tail parameter; the non-smooth prior complicates the computation. The multi-period formulation in [16] is for extreme value distributions with finite tail.

There seems to be no prior work in optimal multi-period estimation of long tail models. This paper addresses the gap. Its contributions are in (i) multi-period models for trends of Pareto or exponential tail parameters based on conjugate priors and (ii) an efficient method for optimal estimation of these models. For the formulated model, the optimal Bayesian estimation of the tail parameter trends separates into two convex optimization problems for tail shape and tail quantile level parameters. These problems can be efficiently solved.

The paper proposes selecting the prior hyper-parameters such that desired smoothing action of the estimator is attained. The method is applied to the estimation of the trends of 10-year extreme event risks in two examples: (i) daily loss in S&P 500 index data, and (ii) peak load in electrical utility data. Since such events do not happen every year, these are non-trivial problems.

II. SINGLE TIME PERIOD PROBLEM

To introduce the problem, we start from a well-studied baseline case of extreme event modeling from the data set

$$D = \{x_i\}_{i=1}^{N}, \tag{1}$$

where scalars x_i are i.i.d. data, i is the sample index, and N is the total number of samples. We consider the right tail events.

This paper models the probability of extreme quantiles. Based on the EVT, we assume that the tail density of the generating distribution for $\{x_i\}$ follows either a Pareto or an exponential distribution. To model the Pareto right tail we use the POT method of EVT for log exceedances $y_i = \log(x_i) - A$, where A is the threshold. For the exponential right tail, we assume the POT exceedances $y_i = x_i - A$.

Using the conditional probability definition, see [17], yields

$$\mathbf{P}(y > u) = \mathbf{P}(y > u | y > 0) \cdot \mathbf{P}(y > 0), \tag{2}$$

where u is a positive parameter, $q = \mathbf{P}(y > 0)$ is the tail quantile level, and $F_{tail}(u) = \mathbf{P}(y > u | y > 0)$ is the tail probability (survival function). In what follows, we estimate q and the tail probability density function $p_{tail}(u)$, which is the derivative of the cumulative density function, $1 - F_{tail}(u)$.

To estimate the tail probability density, we assume that

$$p_{tail}(y) = p_{Exp}(y) = \theta e^{-\theta y}, \tag{3}$$

where $y = \log(x) - A$ for the Pareto tail and $y = x - A$ for the exponential tail. The maximum likelihood estimate (MLE) of the tail rate θ is

$$\hat{\theta} = \arg \max_{\theta} \sum_{y_i \geq 0} (\log \theta - \theta y_i) = \frac{1}{\text{mean}_{y_i \geq 0}\{y_i\}}. \tag{4}$$

For Pareto tail and $\min_{y_i \geq 0}\{y_i\} = 0$, expression (4) is the same as the Hill's estimator [5].

The quantile level parameter q can be estimated from the number of tail points $n = \text{card}\{y_i \geq 0\}$ in the N-point data

set (1). Each data point (1) follows a Bernoulli distribution: it either is in the tail with probability q or not, with probability $1-q$. Thus, n follows a binomial distribution $B(N,q)$, see [18]. For known n, the MLE estimate of the distribution parameter q is (see the Supplementary Material)

$$\hat{q} = \arg \max_{q\in(0,1)} n\log q + (N-n)\log(1-q) = n/N. \quad (5)$$

III. Multi-Period Optimal Estimation Problem

The main contribution of this paper is a multi-period risk trend model extending the formulation of Section II.

A. Problem Formulation

Consider a multi-period dataset, a generalization of (1),

$$\mathcal{D} = \{\{x_{tj}\}_{j=1}^{N_t}\}_{t=1}^{T}, \quad (6)$$

where t is the time period, j is the sample number inside the time period, N_t is number of data points in the time period, and T is the total number of time periods. As in Section II, we analyze log exceedances $y_{tj} = \log x_{tj} - A$ for Pareto tail or $y_{tj} = x_{tj} - A$ for exponential tail. The choice of the threshold A is discussed below in Sections V and VI.

For each time period t, we assume an exponential distribution (3) for y_{tj}, with rate θ_t. For $n_t = \text{card}\{y_{tj} \geq 0\}$, we assume a Binomial distribution (5) with parameter q_t,

$$y_{tj}|(y_{tj} \geq 0, \theta_t) \sim Exp(\theta_t), \quad (7)$$
$$n_t|q_t \sim B(N_t, q_t). \quad (8)$$

In this multi-period setting, the right tail is defined by $\{q_t, \theta_t\}_{t=1}^{T}$. If there are not enough extreme value (POT) data for accurate estimation of the tail model at each time period, a usual approach is to pool all the data and estimate one model for all time periods. This paper proposes a more consistent non-parametric approach. We introduce Bayesian priors for q_t and θ_t and compute maximum a posteriori (MAP) estimates to get their smoothed trends.

B. Prior Distributions

For the trend of the rate parameters θ_t, we assume the prior

$$\theta_t/\theta_{t-1} \sim \text{Gamma}(\beta+1, \beta), \quad (9)$$

where the Gamma distribution $\text{Gamma}(\alpha,\beta)$ is the conjugate prior for the exponential distribution (7). In (9), we set $\alpha = \beta+1$. The mode of $\text{Gamma}(\beta+1,\beta)$ is unity for any $\beta > 0$. If there are no POT data in the current time period ($n_t = 0$), the 'prior-only' MAP yields $\theta_t = \theta_{t-1}$.

The parameter β defines the 'tightness' of the prior (9). In case $\beta = 0$, there is no prior and tail model for each time period is estimated separately. For $\beta = \infty$, the estimates are forced to be the same for all t. This is equivalent to pooling all the data and estimating one model. For $0 < \beta < \infty$, the estimate is smoothed, and more smoothing for larger β.

For the quantile level parameters q_t, we assume the prior

$$\log q_t/\log q_{t-1} = -\log z_t,\ z_t \sim \text{Beta}(\eta, \eta(e-1)+1), \quad (10)$$

where $\text{Beta}(\alpha,\beta)$ is the Beta distribution. The properties of prior (10) are similar to (9). For $\eta = 0$, each q_t is estimated separately. For $\eta = \infty$, a common estimate for all q_t is forced.

The priors (9) and (10) are based on the conjugate priors for the distributions (7) and (8). They allow simulating Markov chains of θ_t, q_t that follow the prior model; this property holds in most filtering models used in signal processing. The prior hyper-parameters β in (9) and η in (10) can be conveniently tuned; this is discussed in Section IV. We next show that priors (9), (10) lead to MAP problems that can be solved efficiently. Section V shows that such MAP estimates provide superior accuracy compared to the baseline methods.

C. Smoothing Filter Formulation

The joint MAP estimation of the formulated Bayesian model (7), (8), (9), (10) splits into two independent problems for θ_t and q_t. These smoothing filter problems are discussed below.

1) Tail Rate Estimation: By combining the likelihood for (7) and the prior (9) we get the posterior. The log posterior for MAP estimation of θ_t has the form

$$L_\theta = \sum_{t=1}^{T} n_t\left(\log\theta_t - \bar{y}_t\theta_t\right) + \beta\sum_{t=2}^{T}\left[\log\frac{\theta_t}{\theta_{t-1}} - \frac{\theta_t}{\theta_{t-1}}\right], \quad (11)$$

where $\bar{y}_t = \text{mean}_{y_{jt}\geq 0}\{y_{jt}\}$.

The MAP estimate of θ_t maximizes L_θ subject to $\theta_t \geq 0$, $(t = 1,\ldots,T)$. After the variable change, $r_t = \log\theta_t$, the constrained non-convex problem of maximizing L_θ (11) becomes an unconstrained convex problem

$$\text{minimize } \sum_{t=1}^{T} n_t\left(\bar{y}_t e^{r_t} - r_t\right) + \beta\sum_{t=2}^{T}\left(e^{\Delta r_t} - \Delta r_t\right), \quad (12)$$

where $\Delta r_t = r_t - r_{t-1}$. The smoothing filter estimate (12) can be found using convex optimization methods, see [19].

2) Tail Quantile Level Estimation: Combining the likelihood based on (8) and the prior (10) gives the log posterior for MAP estimation of q_t in the form

$$L_q = \sum_{t=1}^{T}\left[n_t\log q_t + (N_t - n_t)\log(1-q_t)\right] - \eta\sum_{t=2}^{T}\phi\left(\log\left(\log q_t/\log q_{t-1}\right)\right), \quad (13)$$

$$\phi(x) = e^x - (e-1)\log\left(1-\exp\left(-e^x\right)\right). \quad (14)$$

The MAP estimate of q_t maximizes L_q subject to $0 \leq q_t \leq 1$, $(t = 1,\ldots,T)$. Variable change $u_t = \log(-\log q_t)$, transforms the MAP estimation problem into an unconstrained convex optimization problem

$$\text{minimize } \sum_{t=1}^{T}\left[n_t e^{u_t} - (N_t - n_t)\log\left(1-\exp(-e^{u_t})\right)\right] + \eta\sum_{t=2}^{T}\phi\left(u_t - u_{t-1}\right). \quad (15)$$

D. Risk Trend

The estimated tail model θ_t, q_t allows to trend the risk R_t of exceeding a given threshold u in (2).

$$R_t \quad = \quad \mathbf{P}(y > u|\theta_t, q_t) = q_t \cdot e^{-\theta_t u}. \tag{16}$$

The examples of Section VI set u such that exceeding it is a 1-in-10-years event for pooled data for all time periods.

The model also allows to trend Value at Risk (VaR), see [4]. VaR at confidence level α is defined by $\alpha = \mathbf{P}(x \le \mathrm{VaR}_\alpha)$. For Pareto tail, $y = \log x - A$; using (2), (3) we get

$$\mathrm{VaR}_{\alpha,t} = \exp\left(A - \theta_t^{-1} \log \frac{1-\alpha}{q_t}\right). \tag{17}$$

For exponential tail, $y = x - A$; the expression for $\mathrm{VaR}_{\alpha,t}$ is as in (17) but without the exp.

IV. Smoothing Filter Solution

This section uses quadratic Taylor expansions of (12), (15).

A. Quadratic Approximation

For $r_t = r_{*,t} + v_t$ in (12), the quadratic expansion is

$$\underset{v_t}{\text{minimize}} \sum_{t=1}^{T} a_t \left(v_t - c_t\right)^2 + \sum_{t=2}^{T} b_t \left(v_t - v_{t-1} - d_t\right)^2, \tag{18}$$

where the coefficients depend on the expansion center $r_{*,t}$

$$\begin{aligned} a_t &= n_t \bar{y}_t e^{r_{*,t}}, \qquad & b_t &= \beta e^{r_{*,t} - r_{*,t-1}}, \\ c_t &= (\bar{y}_t e^{r_{*,t}})^{-1} - 1, & d_t &= e^{r_{*,t-1} - r_{*,t}} - 1. \end{aligned} \tag{19}$$

Expanding (15) at $u_t = u_{*,t} + v_t$ yields (18) with

$$\begin{aligned} a_t &= w_{*,t} \left(n_t + q_{*,t} \frac{q_{*,t} - 1 + w_{*,t}}{(1-q_{*,t})^2} (N_t - n_t)\right), \\ b_t &= \eta \phi'' \left(u_{*,t} - u_{*,t-1}\right), \\ c_t &= w_{*,t} a_t^{-1} \left(q_{*,t} (1-q_{*,t})^{-1} (N_t - n_t) - n_t\right), \\ d_t &= -\phi' \left(u_{*,t} - u_{*,t-1}\right) / \phi'' \left(u_{*,t} - u_{*,t-1}\right), \end{aligned} \tag{20}$$

where $w_{*,t} = \exp(u_{*,t})$, $q_{*,t} = \exp(-w_{*,t})$, and the first and second derivative of ϕ (14) are denoted by ϕ' and ϕ''.

B. Almost Uniform Periods

For the tail rate smoothing filter, assume that data in (12) vary little for different time periods, $|\bar{y}_t - \bar{y}_*| \ll 1$, $n_t = n_*$, for $t = 1, \ldots, T$. Then the quadratic expansion (18), (19) centered at the mean solution $r_{*,t} = -\log \bar{y}_*$ has the form

$$\underset{v_t}{\text{minimize}} \sum_{t=1}^{T} \left(v_t - \omega_t\right)^2 + \rho \sum_{t=2}^{T} \left(v_t - v_{t-1}\right)^2, \tag{21}$$

$$\rho = \beta / n_*, \quad \omega_t = 1 - \bar{y}_t / \bar{y}_*. \tag{22}$$

For smoothing filter (15), assume that $|n_t - n_*| \ll n_*$, $N_t = N_*$, for $t = 1, \ldots, T$. The solution for $n_t = n_*$ is $u_t = \log(-\log q_*)$, where $q_* = n_*/N_*$. The quadratic expansion (18), (20), centered at this solution has the form (21) with

$$\rho = \frac{\eta e}{a_*(e-1)}, \quad \omega_t = (n_* - n_t) a_*^{-1} \log q_* \cdot (1-q_*)^{-1}, \tag{23}$$

$$a_* = -\log q_* \left(n_* + q_* (N_* - n_*) \frac{q_* - 1 - \log q_*}{(1-q_*)^2}\right). \tag{24}$$

C. Filter Tuning

The smoothing filter (21) has a single parameter, ρ. It can be expressed through the width w of the non-causal smoothing filter window as (see the Supplementary Material)

$$\rho = e^{-1/w} (e^{-1/w} - 1)^{-2}. \tag{25}$$

This is derived by looking at poles of the filter impulse response, see [20], [21], [22], [23]. Substituting β/n_* in (22) for ρ in (25) yields

$$\beta \quad = \quad n_* \cdot e^{-1/w} \left(e^{-1/w} - 1\right)^{-2}. \tag{26}$$

By substituting ρ from (23), (24) into (25), we get

$$\eta \quad = \quad a_* (1 - 1/e) \cdot e^{-1/w} \left(e^{-1/w} - 1\right)^{-2}. \tag{27}$$

The smoothing filters (12) and (15) can be tuned using rules (26) and (27), (24). For a given data set, start by calculating

$$N_* = \mathrm{mean}_t\{N_t\}, \; n_* = \mathrm{mean}_t\{n_t\}, \; \bar{y}_* = \mathrm{mean}_{y_{tj}>0}\{y_{tj}\}.$$

The prior parameters β and η are then computed based on N_*, n_*, $\bar{y}_*$ in accordance with (26) and (27).

There is a single tuning parameter w. For $w \to 0$, there is no prior; the case $w \to \infty$ means pooling all the data together.

D. Computing the Solution

The smoothing filter solutions $\hat{r}_t$ (12) and $\hat{u}_t$ (15) can be found by the Newton's iterations. Each iteration solves (18), (19) and (18), (20), respectively, using a sparse Hessian matrix.

The Bayesian estimates $\hat{r}_t$ and $\hat{u}_t$ can be assessed by building credible intervals computed through the diagonal elements of the inverse Hessian matrix at the solution, see [24]. The Hessian is defined by a_t and b_t in the quadratic expansion (18). See the Supplementary Material for more detail.

V. Verification and Method Performance

The proposed method was verified in a Monte Carlo simulation study. Each simulation generated a chain of tail rate parameters θ_t for $t = 1, \ldots, T$ using the stochastic model (9) and a chain of quantile level parameters q_t using (10). For each time period t, stochastic models (7), (8) with θ_t and q_t were used to generate n_t tail data points y_{tj}. The fixed parameters of the simulations were: $T = 500$, $N = 10000$, $w = 10$, $N_{runs} = 100$. We used $n_* = \{10, 20, \ldots, 100\}$ and $\bar{y}_* = \{0.1, 1, 10, 100\}$. For each combination $(n_*, \bar{y}_*)$, the hyper-parameters β, η were set using (26), (27); then, N_{runs} simulations were run. The simulated data was processed by smoothing filters (12), (15) to estimate the parameters $\hat{\theta}_t$, $\hat{q}_t$.

Figure 1 illustrates the performance of the proposed method. The root mean square (RMS) estimation error $\theta_t - \hat{\theta}_t$ was averaged over the simulation runs for given $\bar{y}_*$ (upper and lower plots) and n_* (plot argument). The filter used the same value of β as β_{S} employed in the simulation. The proposed method is superior to the baseline methods of pooling all the data ($\beta = \infty$) or looking at each time period separately ($\beta = 0$). It improves the RMS error by factor of 3.4 or more compared to the baseline methods. Similar results hold for $\bar{y}_* = 0.1$ or $\bar{y}_* = 100$ (see the Supplementary Material).

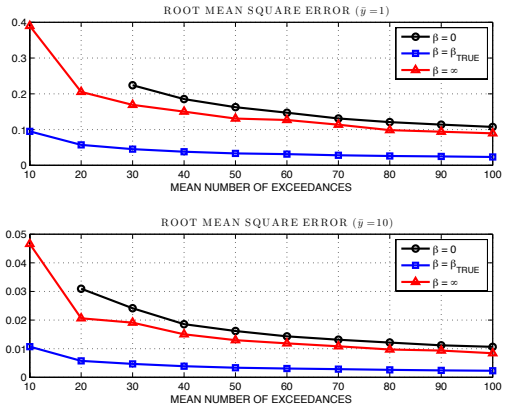

Fig. 1. RMS error of the tail rate estimate: $\bar{y}_* = 1$ (upper plot) and $\bar{y}_* = 10$ (lower plot)

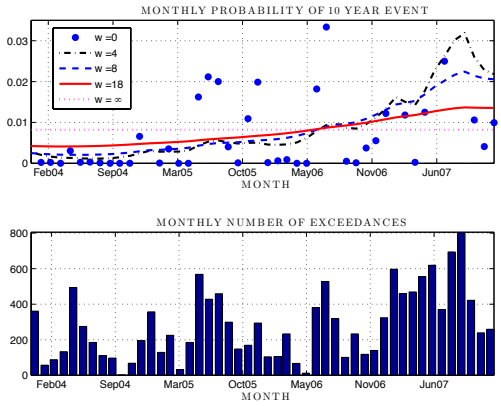

Fig. 2. Probability of a 1-in-10-years power load peak occurring in a given month (upper plot). Exceedance number for each month (lower plot).

VI. Examples

We applied our method to two important signal processing problems: computing risk for financial time series see [4], and power grid monitoring, see [25]. In both examples, we used the tuning rules (26)–(27) with $w = \{0, 4, 8, 18, \infty\}$ to set up the parameters β and η. The tail estimation method parameters used in these examples are summarized in Table I for $w = 18$. The tuning parameter w should be chosen by the analyst.

TABLE I
Parameter Summary ($w = 18$)

	β	η	A	R_*
Power Load	91,196	899,380	1.8	$0.1/8760$
S&P 500 Index	8,340.9	30,517	0.0098	$0.1/252$

A. Peak Power Load

The data set from [26] includes hourly electrical power loads and ambient temperature data for a US utility. The data allow to compute the aggregate load x_t for the utility; it has the range of 0.8 to 3.2 GW with the average of 1.6 GW.

The data covers a time range of 48 consecutive months with the sampling interval of one hour, $N = 38,070$ samples in all. We considered each month as a time period, $T = 48$ periods in all. The power load has strong seasonal component. To reduce its impact, we normalized the load values L_{tj} for month t by the geometric mean value $\bar{L}_m$ for calendar month $m = \mod(t, 12)$ to compute the data (6) as

$$x_{tj} = L_{tj}/\bar{L}_m, \quad (t = 1, \dots, 48). \tag{28}$$

The data has Pareto tail. Table I shows the threshold A used to compute the log-exceedances y_{tj} that have the average exceedance probability $R_* = 0.1/8760$. This is the probability of seeing a 1-in-10-years event in one year of hourly data. The upper plot in Figure 2 compares R_* with $1-(1-R_t)^{720}$, where R_t is from (16) and 720 is number of hours in a month. The monthly threshold exceedance numbers n_t are illustrated by the lower plot of Figure 2. The plots are obtained for the five displayed values of the smoothing parameter w in (26), (27).

B. S&P 500 Risk

We used S&P 500 daily data [27] for a period from 1950 to 2013. Let S_d be the closing index value for trading day d. The relative loss at trading day d is computed as $x_d = S_{d-1}/S_d$. The loss data for each year were combined to create a multi-period data set (6) for $T = 64$ yearly periods.

Table I shows threshold A used to compute the log-exceedances y_{tj} for the Pareto tail. The average risk $R_* = 1-\alpha$ in Table I corresponds to 1-in-10-years loss event and gives confidence level $\alpha = 99.96\%$. We also consider percentage loss $(x_d - 1) \cdot 100\%$. The VaR for the S&P 500 portfolio percentage loss is then $\mathrm{pVaR}_{\alpha,t} = (\mathrm{VaR}_{\alpha,t} - 1)\cdot 100\%$, where $\mathrm{VaR}_{\alpha,t}$ is computed for x_d in accordance with (17). The upper plot in Figure 3 shows $\mathrm{pVaR}_{\alpha,t}$.

The daily loss risk shows an increasing trend, from about 4% in 1950 to to about 6% in 2013. The labels for smoothing parameters w used in (26), (27) are shown in the figure. The lower plot in Figure 3 shows number n_t of the yearly exceedances. See Supplementaty Material for more results.

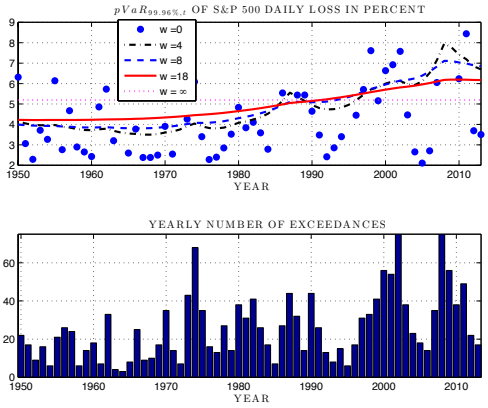

Fig. 3. $\mathrm{pVaR}_{\alpha,t}$ at confidence level $\alpha = 99.96\%$ for S&P loss occurring in a given year (upper plot). Exceedance number for each year (lower plot).

REFERENCES

[1] O. Cappé, E. Moulines, J.-C. Pesquet, A. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *Signal Processing Magazine, IEEE*, vol. 19, no. 3, pp. 14–27, 2002.

[2] S. J. Roberts, "Novelty detection using extreme value statistics," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 146, no. 3, pp. 124–129, 1999.

[3] P. Tsakalides, P. Reveliotis, and C. Nikias, "Scalar quantisation of heavy-tailed signals," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 147, no. 5, pp. 475–484, 2000.

[4] D. Johnston and P. M. Djuric, "The science behind risk management," *Signal Processing Magazine, IEEE*, vol. 28, no. 5, pp. 26–36, 2011.

[5] J. Beirlant, P. Vynckier, and J. L. Teugels, "Tail index estimation, Pareto quantile plots, and regression diagnostics," *Journal of the American Statistical Association*, vol. 91, pp. 1659–1667, December 1996.

[6] J. Danielsson and C. G. Vries, "Beyond the sample: Extreme quantile and probability estimation," tech. rep., Tinbergen Institute Discussion Paper Series, 1997.

[7] R. L. Smith, "Estimating tails of probability distributions," *The Annals of Statistics*, pp. 1174–1207, 1987.

[8] P. Hall and I. Weissman, "On the estimation of extreme tail probabilities," *The Annals of Statistics*, pp. 1311–1326, 1997.

[9] B. C. Arnold and S. J. Press, "Bayesian estimation and prediction for Pareto data," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 1079–1084, 1989.

[10] J. Diebolt, M.-A. El-Aroui, M. Garrido, and S. Girard, "Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling," *Extremes*, vol. 8, no. 1-2, pp. 57–78, 2005.

[11] Y. Chen, P. Breen, and N. Andrew, "Impacts of outliers and mis-specification of priors on Bayesian fisheries-stock assessment," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 57, no. 11, pp. 2293–2305, 2000.

[12] C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns, "Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles," *Journal of Climate*, vol. 18, no. 10, pp. 1524–1540, 2005.

[13] R. L. Smith, "Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone," *Statistical Science*, pp. 367–377, 1989.

[14] R. L. Smith and T. S. Shively, "Point process approach to modeling trends in tropospheric ozone based on exceedances of a high threshold," *Atmospheric Environment*, vol. 29, no. 23, pp. 3489–3499, 1995.

[15] V. Chavez-Demoulin, P. Embrechts, and S. Sardy, "Extreme-quantile tracking for financial time series," *Journal of Econometrics*, vol. 181, no. 1, pp. 44–52, 2014.

[16] D. A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, "An extreme function theory for novelty detection," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 28–37, 2013.

[17] A. Shiryaev and S. Wilson, *Probability*. Graduate Texts in Mathematics, Springer New York, 1995.

[18] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer, 2007.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[20] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*. New Jersey: Prentice Hall, 4th ed., April 2006.

[21] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing*, vol. 1. Prentice-Hall Signal Processing Series, Englewood Cliffs: Prentice-Hall, 1984.

[22] J. S. Lim, "Two-dimensional signal and image processing," *Englewood Cliffs, NJ, Prentice Hall, 1990, 710 p.*, vol. 1, 1990.

[23] R. C. Gonzales and R. E. Woods, *Digital Image Processing, 2-nd Edition*. Prentice Hall, 2002.

[24] W. Q. Meeker and L. A. Escobar, "Teaching about approximate confidence regions based on maximum likelihood estimation," *The American Statistician*, vol. 49, no. 1, pp. 48–53, 1995.

[25] G. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, and B. Wollenberg, "Monitoring and optimization for power grids: A signal processing perspective," *Signal Processing Magazine, IEEE*, vol. 30, no. 5, pp. 107–128, 2013.

[26] Kaggle.com, "Global energy forecasting competition 2012 load forecasting." Available: https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting.

[27] Quandl.com, "S&P 500 index." Available: http://www.quandl.com/YAHOO/INDEX_GSPC-S-P-500-Index.

[28] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, pp. 54–75, 1986.

APPENDIX

APPENDIX A: BERNOULLI LIKELIHOOD IN (5)

Each point in the data set (1) follows a Bernoulli distribution: it either belongs to the distribution tail with probability q or or not, with probability $1-q$. The probability of having n points belonging to the tail is given by

$$p_B(n|q) = C_N^n q^n (1-q)^{N-n}, \tag{A.1}$$

where $C_N^n = N!/(n!(N-n)!)$. For known n, the Maximum Likelihood Estimate (MLE) of q can be obtained by maximizing $\log p_B(n|q)$ in (A.1). This yields (5)

$$\hat{q} = \arg \max_{q\in(0,1)} n\log q + (N-n)\log(1-q) = n/N.$$

APPENDIX B: QUADRATIC APPROXIMATION (18)

Non-linear filters (12) and (15) minimize loss functions of the form

$$L(x) = \sum_{t=1}^{T} f(x_t) + \sum_{t=2}^{T} g(\Delta x_t), \tag{A.2}$$

where $\Delta x_t = x_t - x_{t-1}$. This expression hides the dependence on parameters and only shows explicit dependence on the decision variables that we optimize over. Quadratically expanding (A.2) in a Taylor series about $x_{*,t}$, with $x_t = x_{*,t} + v_t$ yields, after throwing out the constant terms

$$\begin{aligned} L(x) \approx &\sum_{t=1}^{T}\left(f'(x_{*,t})v_t + \frac{1}{2}f''(x_{*,t})v_t^2\right) \\ &+ \sum_{t=2}^{T}\left(g'(\Delta x_{*,t})\Delta v_t + \frac{1}{2}g''(\Delta x_{*,t})(\Delta v_t)^2\right), \end{aligned} \tag{A.3}$$

where $\Delta v_t = v_t - v_{t-1}$. We denote the first and second derivative of a function f with f' and f'' respectively and introduce the notation

$$\begin{aligned} &a_t = f''(x_{*,t}), \quad c_t = -f'(x_{*,t})/f''(x_{*,t}), \\ &b_t = g''(\Delta x_{*,t}),\ d_t = -g'(\Delta x_{*,t})/g''(\Delta x_{*,t}). \end{aligned} \tag{A.4}$$

Then minimization of (A.3) can be written in the form (18)

$$\underset{v_t}{\text{minimize}} \sum_{t=1}^{T} a_t\left(v_t - c_t\right)^2 + \sum_{t=2}^{T} b_t\left(v_t - v_{t-1} - d_t\right)^2.$$

APPENDIX C: TAIL RATE FILTER FORMULAS (19), (22)

Consider tail rate filter (12). It minimizes loss function of the form (A.2) where

$$f(r_t) = n_t\left(\bar{y}_t e^{r_t} - r_t\right), \quad g(\Delta r_t) = \beta\left(e^{\Delta r_t} - \Delta r_t\right).$$

By substituting these expressions into (A.4) we compute a_t, b_t, c_t, d_t to get (19)

$$\begin{aligned} &a_t = n_t \bar{y}_t e^{r_{*,t}}, \qquad b_t = \beta e^{r_{*,t} - r_{*,t-1}}, \\ &c_t = \left(\bar{y}_t e^{r_{*,t}}\right)^{-1} - 1,\ d_t = -1 + e^{r_{*,t-1} - r_{*,t}}. \end{aligned}$$

To understand performance of the smoothing filter for the tail rate, assume that data in (12) vary little for different time periods, $|\bar{y}_t - \bar{y}_*| \ll 1$, $n_t = n_*$, for $t = 1, \ldots, T$.

First, consider a baseline solution where $\bar{y}_t = \bar{y}_*$ for $t = 1, \ldots, T$. In that case, solving (12) yields $r_t = r_*$, where $\bar{y}_* e^{r_*} = 1$.

Quadratic expansion (18) centered at this baseline solution has coefficients (19) where

$$\begin{aligned} d_t &= 0, \quad b_t = \beta, \quad a_t = n_*, &\text{(A.5)} \\ c_t &= \left(\bar{y}_t e^{r_{*,t}}\right)^{-1} - 1 \approx \frac{1}{\bar{y}_* e^{r_*}\left(1 + (\bar{y}_t - \bar{y}_*)/\bar{y}_*\right)} - 1 \\ &\approx -\frac{\bar{y}_t - \bar{y}_*}{\bar{y}_*} = 1 - \bar{y}_t/\bar{y}_*. &\text{(A.6)} \end{aligned}$$

By dividing (18) through by $a_t = n_*$ and using (A.5), (A.6) we get filter (21) with parameters (22)

$$\rho = \beta/n_*, \quad \omega_t = -z_t/\bar{y}_*.$$

APPENDIX D: TAIL QUANTILE LEVEL FILTER FORMULAS (20), (23)

Consider tail quantile level filter (15). It minimizes loss function of the form (A.2), where

$$\begin{aligned} f(u_t) &= n_t e^{u_t} - (N_t - n_t)\log\left(1 - \exp(-e^{u_t})\right), \\ g(\Delta u_t) &= \eta\phi\left(\Delta u_t\right), \end{aligned}$$

where $\phi(x)$ is defined in (14).

By substituting these expressions into (A.4) we compute a_t, b_t, c_t, d_t to get (20)

$$\begin{aligned} &a_t = w_{*,t}\left(n_t + q_{*,t}\frac{q_{*,t} - 1 + w_{*,t}}{(1-q_{*,t})^2}(N_t - n_t)\right), \\ &b_t = \eta\phi''\left(u_{*,t} - u_{*,t-1}\right), \\ &c_t = w_{*,t} a_t^{-1}\left(q_{*,t}(1-q_{*,t})^{-1}(N_t - n_t) - n_t\right), \\ &d_t = -\phi'\left(u_{*,t} - u_{*,t-1}\right)/\phi''\left(u_{*,t} - u_{*,t-1}\right), \end{aligned}$$

where $q_{*,t} = \exp(-\exp(u_{*,t}))$. The first and second derivative of $\phi(x)$ is given by

$$\begin{aligned} \phi'(x) &= e^x\left(1 - \frac{e-1}{\exp\left(e^x\right) - 1}\right), \\ \phi''(x) &= e^x\left(1 - (e-1)\frac{\exp(e^x)\left(1 - e^x\right) - 1}{\left(\exp(e^x) - 1\right)^2}\right). \end{aligned}$$

To understand the performance of the smoothing filter for the tail quantile level, assume that data in (15) vary little for different time periods, $|n_t - n_*| \ll n_*$, $N_t = N_*$, for $t = 1, \ldots, T$.

First, consider a baseline solution, where $n_t = n_*$ for $t = 1, \ldots, T$. In that case, solving (15) yields $u_t = u_*$, where $u_* = \log\log(N_*/n_*)$.

Quadratic expansion (18) centered at this baseline solution has coefficients (19), where $d_t = 0$, $b_t = \eta e/(e-1)$. The expressions for a_t, c_t use $N_t = N_*$ and $q_{*,t} = q_* = \exp(-\exp(u_*)) = n_*/N_*$. We approximate $a_t \approx a_*$ by using $n_t = n_*$ in (20)

$$a_* = -\log q_* \cdot \left(n_* + q_*(N_* - n_*)\frac{q_* - 1 - \log q_*}{(1-q_*)^2}\right). \tag{A.7}$$

Using this approximation of a_t, we get

$$\begin{aligned} c_t &= -\log q_* \left((N_* - n_t) \frac{q_*}{1-q_*} - n_t \right) / a_* \\ &= (n_* - n_t) a_*^{-1} \log q_* \cdot (1-q_*)^{-1}. \end{aligned} \tag{A.8}$$

Dividing (18) through by a_* in (A.7) and using (A.8) we get filter (21) with parameters (22)

$$\begin{aligned} \rho &= \eta e a_*^{-1} (e-1)^{-1}, \\ \omega_t &= (n_* - n_t) a_*^{-1} \log q_* \cdot (1-q_*)^{-1}. \end{aligned}$$

Appendix E: Newton's Method in Section IV-D

Subsection IV-D mentions solving the optimization problem with the Newton's method. Each Newton's step optimizes the quadratic expansion of the form (18)

$$\text{minimize } (v-c)^T Q (v-c) + (Dv-d)^T B (Dv-d),$$

$$Q = \operatorname{diag}(a_1, ..., a_t, ..., a_T), \tag{A.9}$$

$$B = \operatorname{diag}(b_1, ..., b_t, ..., b_T), \tag{A.10}$$

$$c = [c_1, ..., c_t, ..., c_T]^T, \tag{A.11}$$

$$d = [d_1, ..., d_t, ..., d_T]^T, \tag{A.12}$$

$$v = [v_1, ..., v_t, ..., v_T]^T, \tag{A.13}$$

where D is the two-diagonal first difference matrix with -1 on the main diagonal (except the last, zero, entry) and 1 above the main diagonal. Matrices Q, B, and vectors c, d, depend on the expansion center $x_* = [x_{*,1}, ..., x_{*,t}, ..., x_{*,T}]^T$. For the tail rate and tail quantile level parameter filters, the entries of these matrices are described by (19) and (20) respectively.

Differentiating the optimization index with respect to vector v yields a system of linear equations. Solving it for v gives

$$v = H^{-1} \left(Qc + D^T B d \right), \tag{A.14}$$

$$H = Q + D^T B D, \tag{A.15}$$

where Q, B, c, d, are given by (A.9)–(A.12).

The Newton's method iterations go as follows. Let $x^{(n)}$ be the approximate solution at iteration n. We compute matrices (A.9)–(A.12) using $x_* = x^{(n)}$ as the approximation center. Then, the Newton's step $v^{(n)}$ is computed from (A.14). It is used to get the next iteration of the approximate solution as

$$x_t^{(n+1)} \leftarrow x_t^{(n)} + v_t^{(n)}. \tag{A.16}$$

The iterations continue until convergence is achieved. Since the problem is convex and smooth, the Newton's method iterations are guaranteed to converge. In the examples, they converge very fast.

Appendix F: Credible Intervals in Section IV-D

Credible intervals can be computed using quadratic approximation of the log posterior. This approach, known as Wald approximation, is equivalent to approximating the posterior as a normal distribution, see [24].

In (12) and (15), the log-posteriors for θ_t and q_t are parametrized by r_t and u_t. Since level surfaces are invariant under re-parametrizations, see [24], the credible intervals for θ_t and q_t can be found through the intervals for r_t and u_t.

Consider a quadratic expansion of $L(x)$, where $L(x)$ represents either $L_r(r)$ or $L_u(u)$, around the optimal solution $x - x_*$. The gradient at the optimum $(dL/dx)|_{x=x_*} = 0$, thus,

$$L(x) \approx L(x_*) + \frac{1}{2} (x - x_*)^T H_* (x - x_*), \tag{A.17}$$

where H_* is the Hessian (A.15) computed at the optimum x_*.

In the Wald approximation, the Fisher information matrix H_* is the inverse of the covariance matrix of x, see [24]. For any constant vector s, the linear combination $s^T x$ of the components of x follows the approximate normal distribution

$$s^T x \sim \mathcal{N}(s^T x_*, s^T H_*^{-1} s) \tag{A.18}$$

The credible interval for x_k is computed by taking $s_k = 1$ and $s_j = 0$ for all other j. It is the confidence interval for the univariate normal distribution (A.18), see [28]. Using (A.15), (A.18), we get the intervals $[r_{-,t}, r_{+,t}]$ and $[u_{-,t}, u_{+,t}]$ as

$$r_{\pm,t} = r_{*,t} \pm c \sqrt{\left(H_*^{-1}\right)_{tt}}, \tag{A.19}$$

$$u_{\pm,t} = u_{*,t} \pm c \sqrt{\left(H_*^{-1}\right)_{tt}}, \tag{A.20}$$

where constant c is the standard normal quantile at a given confidence level. For 95% confidence level $c = 1.96$. The credible intervals for $[\theta_{-,t}, \theta_{+,t}]$ and $[q_{-,t}, q_{+,t}]$ are then

$$\theta_{\pm,t} = \exp(r_{\pm,t}), \quad q_{\pm,t} = \exp(-e^{u_{\mp,t}}).$$

Appendix G: Smoothing Window Width in (25)

Consider a linear filter of the form (21) given by

$$\text{minimize } \sum_{t=1}^{T} (x_t - y_t)^2 + \rho \sum_{t=2}^{T} (x_t - x_{t-1})^2, \tag{A.21}$$

where x_t is the input sequence and y_t is the output sequence While (A.21) implies finite sequences of length T, we further assume $T \to \infty$ and consider x_t and y_t as time series. By differentiating with respect to x_t and introducing a unit time shift operator z we get

$$\left(1 + 2\rho - \rho z^{-1} - \rho z\right) x_t = y_t. \tag{A.22}$$

This expression describes a non-causal i.i.r. filter. Its poles can be obtained by factorizing the polynomial in the l.h.s. of (A.22) into the stable and anti-stable factors of the form $(z-\delta)(z^{-1}-\delta)$. From (A.22) we get

$$\delta = (1 + 2\rho - \sqrt{1+4\rho})/(2\rho), \tag{A.23}$$

where $\delta < 1$. Both the causal and the anti-causal part of the impulse response decay as δ^k, where k is the distance from the response center. We define filter impulse response width w from $\delta^w = e^{-1}$. This yields $\delta = e^{-1/w}$. By substituting expression (A.23) for δ and solving for ρ we get (25)

$$\rho = e^{-1/w} (e^{-1/w} - 1)^{-2}.$$

APPENDIX H: ADDITIONAL VERIFICATION FOR SECTION V

Figure 1 in Section V shows Monte Carlo simulation results for the proposed method obtained for $\bar{y}_* = 1$ and 10.

Additional results obtained for $\bar{y}_* = 0.1$ and 100 are shown in Figure 4. The top and lower plot in Figure 4 shows the case of $\bar{y}_* = 0.1, 100$ respectively. In both plots, the RMS error results for $\beta = \beta_{TRUE}$ are better than for $\beta = 0$ and $\beta = \infty$. This indicates that the method developed is the best by good margin for all studied tail length parameters $\bar{y}_*$.

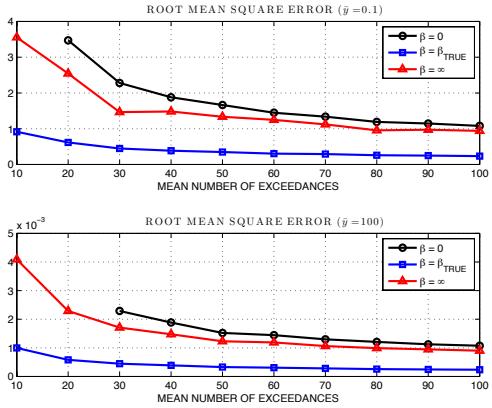

Fig. 4. Additional results of the Monte Carlo verification study for $\bar{y}_* = 0.1$ (the upper plot) and $\bar{y}_* = 100$ (the lower plot).

APPENDIX I: ADDITIONAL RESULTS FOR SECTION VI

In the examples of Section VI, Figure 2 shows the risk of of 1-in-10-years power peak. The risk (16) is based on the tail parameter estimates obtained by the proposed filters. These tail rate and quantile level parameters are estimated for each monthly time period in the energy load dataset. Figure 5 shows the Bayesian estimate trends of θ_t and q_t for five values of w as displayed in the plot. For $w = 18$, Figure 5 shows the 95% credible intervals computed as discussed in Appendix F.

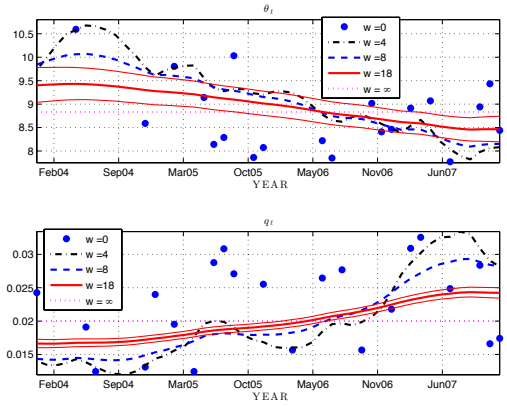

Fig. 5. Tail rate θ_t and tail quantile level q_t parameters for the power load data with 95% credible intervals marked as thin red lines.

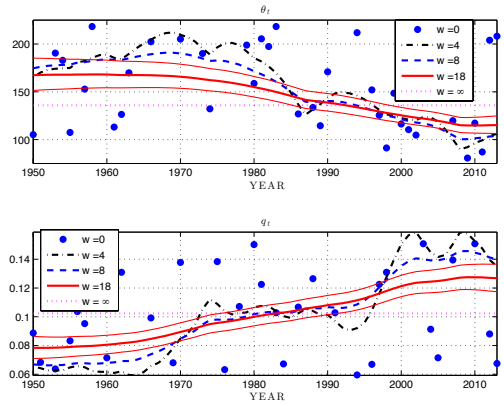

Fig. 6. Tail rate θ_t and tail quantile level q_t parameters for the S&P 500 index with 95% credible intervals marked as thin red lines.

Figure 3 shows VaR (17) for the S&P 500 index based on the tail parameters estimated by the filters. Figure 6 shows the trends of tail rate θ_t and quantile level q_t estimated for the S&P 500 index. For the $w = 18$ curve, Figure 6 shows the 95% credible intervals.

The trends in both Figure 5 and Figure 6 indicate that the data is increasingly heavier tailed with time. Two factors contribute to that. First, the tail rate parameter θ_t decreases over time. Since tail rate θ_t is effectively the inverse tail length for time period t, this means the tail is becoming longer. Second, the tail quantile level parameter q_t is increasing with time. This means the probability of an exceedance occurring in time period t is increasing. The risk trend in Figure 2 and pVaR trend in Figure 3 reflect that the tails in these two problems are becoming increasingly heavy with time.

The upper plot in Figure 7 shows the trend of risk probability $1 - (1 - R_t)^{252}$, with R_t from (16) and 252 trading days in a year. The lower plot shows the worst yearly log returns (the minimum $\log x_d$ in Section VI-B) for each time period. This plot confirms that the loss risk increases over the time.

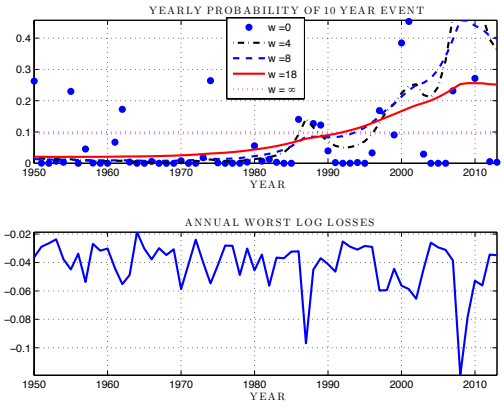

Fig. 7. Plots of risk trend of S&P 500 index losses and log of worst relative yearly returns.