Statistical methods for modeling of multidimensional systems

Julie Bessac (Argonne National Laboratory)

April 2022

Outline

Part 1: Context and workflow of statistical modeling

Part 2: Statistical metrics to assess lossy compressibility of scientific datasets

Part 3: Non-stationary bulk and tails of temperature

Introduction to statistical modeling

Motivations:

- provide and quantify uncertainty (data, prediction, model, ...)
- comprehensive description of data (correlation, variable importance, extremes, ...)
- overcome lack of data (conditional emulation, prediction, fusion, ...)
- complement physics-driven models
- emulate realistic samples very efficiently

Introduction to statistical modeling

Motivations:

- provide and quantify uncertainty (data, prediction, model, ...)
- comprehensive description of data (correlation, variable importance, extremes, ...)
- overcome lack of data (conditional emulation, prediction, fusion, ...)
- complement physics-driven models
- emulate realistic samples very efficiently

Approach:

$\rightarrow$ Reproduce target quantities of interest
probabilistic distribution, time series dynamics, space-time dependence, interaction between variables, ...

$\rightarrow$ Build parametric structures to describe distributions, covariances, ... $\rightarrow$ our focus

Introduction to statistical modeling

Motivations:

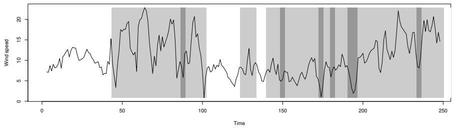
- provide and quantify uncertainty (data, prediction, model, ...)
- comprehensive description of data (correlation, variable importance, extremes, ...)
- overcome lack of data (conditional emulation, prediction, fusion, ...)
- complement physics-driven models
- emulate realistic samples very efficiently

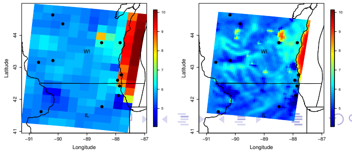
Approach:

→ Reproduce target quantities of interest
probabilistic distribution, time series dynamics, space-time dependence, interaction between variables, ...
→ Build parametric structures to describe distributions, covariances, ... → our focus

Challenges:

nature of the data, amount of data, non-stationarity, dependencies and correlations, multiple scales, rare events, errors and uncertainty in the data





Statistical modeling workflow

Data analysis & problem formulation

$\downarrow$

Construction of a parametric model: (Research topic)

Ex.: Autoregressive process, $X_{t+1} = \rho X_t + \sigma \epsilon_{t+1}, t \geq 0$

$\downarrow$

Estimation of the model parameters: (Research topic)

Least square, Maximum likelihood, ...

Ex.: Estimate ρ and σ to match data as well as possible

$\downarrow$

Evaluation of fitted model: (Research topic)

Associated with the targeted application and features and use of scalar metrics (mean-squared error, scores, ...)

$\downarrow$

Inference, prediction, simulation by Monte-Carlo,... (Research topic)

Ex.:

- Prediction: $\hat{x}_{t+1} = \rho x_t$
- Simulation: $x_0 \sim P_0$, $\epsilon \sim P_\epsilon$, and *for* $(i \; in \; 1 : N)$, $x_{i+1} = \rho x_i + \sigma \epsilon_i$

Exploring lossy compressibility through statistical correlations of scientific datasets

Julie Bessac (ANL), Robert Underwood (ANL), David Krasowska (Clemson University), Sheng Di (ANL), Jon Calhoun (Clemson University), Franck Cappello (ANL)

Krasowska, D., Bessac, J., Calhoun, J., Underwood, R., Di, S., and Cappello, F. (2021). Exploring lossy compressibility through statistical correlations of scientific datasets.
In *7th International Workshop on Data Analysis and Reduction for Big Scientific Data in conjunction with SC '21: The International Conference for High Performance Computing, Networking, Storage and Analysis - https://arxiv.org/pdf/2111.13789.pdf*, pages 47–53

Context and goals

· Lossy compressors are increasingly adopted in scientific research
→ tackle large amount of data generated by experiments or simulations
→ facilitating data storage and movement in high-performance computing systems

Context and goals

· Lossy compressors are increasingly adopted in scientific research
→ tackle large amount of data generated by experiments or simulations
→ facilitating data storage and movement in high-performance computing systems

· In lossless compression, entropy provides theoretical limit on compressibility of data but there are no equivalent for lossy compressors

Characterize statistics of the data that impact lossy compression, e.g. correlation structures, patterns, range of values, spatial heterogeneity, ...
& build prediction models for compression ratios

Context and goals

· Lossy compressors are increasingly adopted in scientific research
→ tackle large amount of data generated by experiments or simulations
→ facilitating data storage and movement in high-performance computing systems

· In lossless compression, entropy provides theoretical limit on compressibility of data but there are no equivalent for lossy compressors

Characterize statistics of the data that impact lossy compression, e.g. correlation structures, patterns, range of values, spatial heterogeneity, ...
& build prediction models for compression ratios

· These models will form the first step towards evaluating theoretical limits of lossy compressibility
→ how far are existing compressors to optimality
→ help optimize compressors allow maximum efficiency for storing scientific datasets

Prediction of lossy compression ratios

Variety of compressors

SZ (prediction-based), ZFP (transform-based), MGARD (multigrid), Digit Rounding & Bit Grooming (rounding-based)
$\rightarrow$ compression ratios (CR)

Prediction of lossy compression ratios

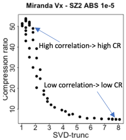
Variety of compressors

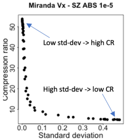
SZ (prediction-based), ZFP (transform-based), MGARD (multigrid), Digit Rounding & Bit Grooming (rounding-based) $\rightarrow$ compression ratios (CR)

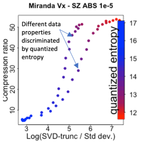
Statistics of interest (compressor-free)

- Correlation strength extracted from singular value decomposition SVD truncation
- Standard deviation (variability and value range)
- Lossyness and patterns from quantized entropy

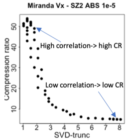
Data used to train regression models numerical simulations (cosmology, atmospheric, hydrodynamic)

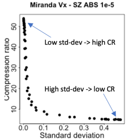


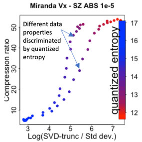




Prediction of lossy compression ratios







Variety of compressors

SZ (prediction-based), ZFP (transform-based), MGARD (multigrid), Digit Rounding & Bit Grooming (rounding-based)
$\rightarrow$ compression ratios (CR)

Statistics of interest (compressor-free)

- Correlation strength extracted from singular value decomposition SVD truncation
- Standard deviation (variability and value range)
- Lossyness and patterns from quantized entropy

Data used to train regression models numerical simulations (cosmology, atmospheric, hydrodynamic)

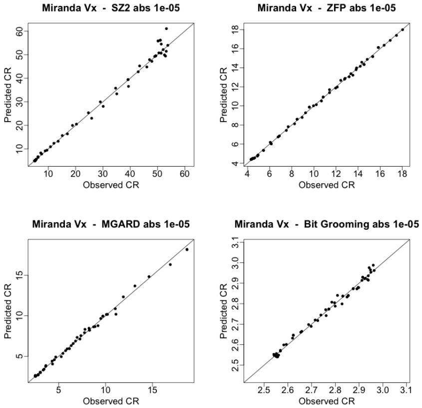
Regression models

$$\log(\text{CR}) = s(\log(\text{q-ent})) + s\left(\log\left(\frac{\text{SVD-trunc}}{\sigma}\right)\right) + ti\left(\log(\text{q-ent}), \log\left(\frac{\text{SVD-trunc}}{\sigma}\right)\right) + \epsilon,$$

$\rightarrow$ regression fitted on observed CR and statistics computed on the data

Results and discussion

Out-of-sample prediction of CR



Very good compression ratio prediction with spline regression

Framework still relies on the use of compressors → how to go further and provide a compressor-free characterization of compressibility?

Interesting questions on the statistical side → how to summarize multiscale and-or correlation heterogeneity into scalar quantities?

Nonstationary seasonal model for daily mean temperature distribution bridging bulk and tails

Mitchell Krock (Rutgers University), Julie Bessac (Argonne National Laboratory), Michael Stein (Rutgers University), Adam Monahan (University of Victoria)

Krock, M., Bessac, J., Stein, M. L., and Monahan, A. (2022). Seasonal bulk-and-tails model with long-term trends for temperature - https://arxiv.org/pdf/2110.10046.pdf.
Weather and Climate Extremes - In Press

Motivations and data

· While global mean temperature has been rising
$\rightarrow$ regional temperature exhibits various patterns of change, including extremes
e.g. warmest temperatures are stretching and main cold temperature are shrinking

· Daily mean surface air temperature (SAT) from
NCEI's Global Surface Summary of the Day

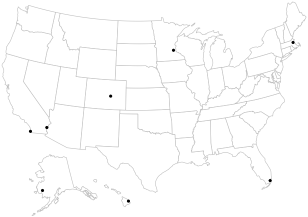
Motivations and data

· While global mean temperature has been rising
→ regional temperature exhibits various patterns of change, including extremes
e.g. warmest temperatures are stretching and main cold temperature are shrinking

· Daily mean surface air temperature (SAT) from NCEI's Global Surface Summary of the Day

· **Objective**: Nonstationary (seasonal and long-term trend) model for **entire distribution** of daily temperature, focusing on behavior in both tails (hot and cold extremes) [Krock et al., 2022]

· Most statistical methods for extremes focus on one tail of the distribution (Generalized Extreme Value distribution, Generalized Pareto distribution)



Eight locations with very different climates and geographies

Building on (Stein 2020) that introduced "Bulk-And-Tails" (BATs) model for the entire distribution with flexible behavior in both tails

$$F_\theta(x) = T_\nu(H_\theta(x)) \text{ with } T_\nu \ t\text{-cdf with } \nu \text{ d.o.f.}$$

$$H_\theta(x) = \left(1 + \kappa_1 \Psi\left(\frac{x - \phi_1}{\tau_1}\right)\right)^{1/\kappa_1} - \left(1 + \kappa_0 \Psi\left(\frac{\phi_0 - x}{\tau_0}\right)\right)^{1/\kappa_0}$$

$$\Psi(x) = \log(1 + \exp(x)) \quad \text{and} \quad \theta = (\underbrace{\kappa_0, \tau_0, \phi_0}_{\text{Lower tail}}, \underbrace{\kappa_1, \tau_1, \phi_1}_{\text{Upper tail}})$$

· Comprehensive modeling of each tail
  → Heaviness: κ ; Location: ϕ ; Spread: τ

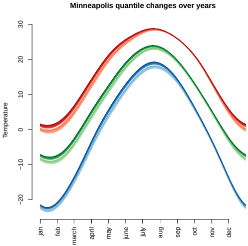
Building on (Stein 2020) that introduced "Bulk-And-Tails" (BATs) model for the entire distribution with flexible behavior in both tails

$$F_\theta(x) = T_\nu(H_\theta(x)) \text{ with } T_\nu \text{ } t\text{-cdf with } \nu \text{ d.o.f.}$$

$$H_\theta(x) = \left(1 + \kappa_1 \Psi\left(\frac{x - \phi_1}{\tau_1}\right)\right)^{1/\kappa_1} - \left(1 + \kappa_0 \Psi\left(\frac{\phi_0 - x}{\tau_0}\right)\right)^{1/\kappa_0}$$

$$\Psi(x) = \log(1 + \exp(x)) \quad \text{and} \quad \theta = (\underbrace{\kappa_0, \tau_0, \phi_0}_{\text{Lower tail}}, \underbrace{\kappa_1, \tau_1, \phi_1}_{\text{Upper tail}})$$

· Comprehensive modeling of each tail
$\rightarrow$ Heaviness: κ ; Location: ϕ ; Spread: τ



Non-stationary seasonal extension [Krock et al., 2022]

Location parameters: $\phi_{\cdot}(day, year) =$ seasonal$(day)+$trend$(year)+$seasonal$(day)\times$trend$(year)$

Scale parameters: $\tau_{\cdot}(day) =$ seasonal(day)

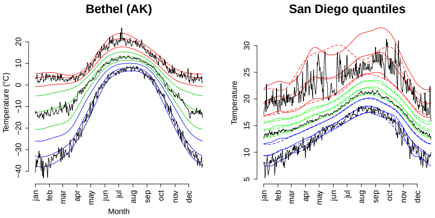
Shape parameters estimated fixed across days and years

· Long-term trend approximated by log(CO2 equivalent) (yearly covariate, proxy for climate change induced by greenhouse gases)

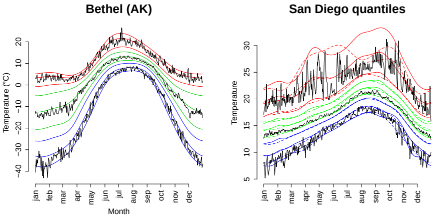
Seasonal quantiles

· BATs **quantiles** for year 2020: 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, and 0.9, 0.99, 0.999

· Black lines: observation daily minimum/median/maximum taken over all years

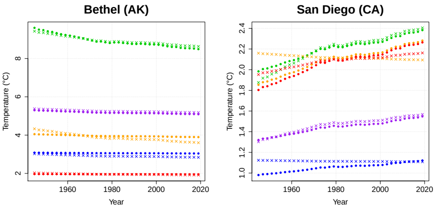


Seasonal quantiles



- BATs **quantiles** for year 2020: 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, and 0.9, 0.99, 0.999
- Black lines: observation daily minimum/median/maximum taken over all years

Year-to-year quantile evaluation



Average differences between quantiles for each year based on the BATs model (o) and quantile regression (x)

$q_{0.99} - q_{0.95}$, $q_{0.95} - q_{0.75}$

$q_{0.75} - q_{0.25}$

$q_{0.05} - q_{0.01}$, $q_{0.25} - q_{0.05}$

References I

Janke, T., Ghanmi, M., and Steinke, F. (2021).
Implicit generative copulas.
Advances in Neural Information Processing Systems, 34.

Krasowska, D., Bessac, J., Calhoun, J., Underwood, R., Di, S., and Cappello, F. (2021).
Exploring lossy compressibility through statistical correlations of scientific datasets.
In *7th International Workshop on Data Analysis and Reduction for Big Scientific Data in conjunction with SC '21: The International Conference for High Performance Computing, Networking, Storage and Analysis - https://arxiv.org/pdf/2111.13789.pdf*, pages 47–53.

Krock, M., Bessac, J., Stein, M. L., and Monahan, A. (2022).
Seasonal bulk-and-tails model with long-term trends for temperature - https://arxiv.org/pdf/2110.10046.pdf.
Weather and Climate Extremes - In Press.