NPL-SAMBA ITT potential projects

Alistair Forbes[1]

[1]National Physical Laboratory, UK
Data Science Group

University of Bath

Outline

NPL National Physical Laboratory

Fitting a model to data

- Standard data fitting model

$$\boldsymbol{y} = C\boldsymbol{a} + \epsilon, \quad \epsilon \in \mathrm{N}(\boldsymbol{0}, \sigma^2 I)$$

- $\boldsymbol{y}$ is an $m \times n$ data vector, $\boldsymbol{a}$ parameters of the model
- C is an $m \times n$ observation matrix, e.g. basis functions evaluated at $\boldsymbol{x}$
- ϵ is an $m \times n$ vector of independent random effects associated with the measuring system
- Least squares model fit

$$\hat{\boldsymbol{a}} = (C^{\mathrm{T}} C)^{-1} C^{\mathrm{T}} \boldsymbol{y} = R_1^{-1} Q_1^{\mathrm{T}} \boldsymbol{y}, \quad C = Q_1 R_1$$

$$\hat{\boldsymbol{y}} = C\hat{\boldsymbol{a}} = C(C^{\mathrm{T}} C)^{-1} C^{\mathrm{T}} \boldsymbol{y} = Q_1 Q_1^{\mathrm{T}} \boldsymbol{y}$$

Effective number of degrees of freedom in a model

- If $\hat{\boldsymbol{y}} = H\boldsymbol{y}$, the sum of the eigenvalues of H is a measure of the number of degrees of freedom associated with the model.
- Least squares model fit

$$\hat{\boldsymbol{y}} = C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}} = Q_1 Q_1^{\mathrm{T}}\boldsymbol{y}$$

- $Q_1 Q_1^{\mathrm{T}}$ is a projection with n eigenvalues equal to 1, all others 0.

Correlated systematic effects

- Extension of the standard model:

$$\boldsymbol{y} = C\boldsymbol{a} + \boldsymbol{e} + \epsilon, \quad \boldsymbol{e} \in \mathrm{N}(\boldsymbol{0}, V_0), \quad \epsilon \in \mathrm{N}(\boldsymbol{0}, \sigma^2 I)$$

NPL
National Physical Laboratory

Gauss Markov regression

- Combined variance matrix, Choleski decomposition

$$V = V_0 + \sigma^2 I = LL^{\mathrm{T}}, \quad \tilde{\boldsymbol{y}} = L^{-1}\boldsymbol{y}, \quad \tilde{C} = L^{-1}C$$

$$\tilde{\boldsymbol{y}} = \tilde{C}\boldsymbol{a} + \tilde{\epsilon}, \quad \tilde{\epsilon} \in \mathrm{N}(\boldsymbol{0}, I)$$

- Effective degrees of freedom: transformed problem

$$\hat{\tilde{\boldsymbol{y}}} = \tilde{Q}_1\tilde{Q}_1^{\mathrm{T}}\tilde{\boldsymbol{y}}$$

- Effective degrees of freedom: original problem

$$\hat{\boldsymbol{y}} = L\hat{\tilde{\boldsymbol{y}}} = L\tilde{Q}_1\tilde{Q}_1^{\mathrm{T}}L^{-1}\boldsymbol{y}$$

NPL National Physical Laboratory

Explicit effects model

- Same extended model

$$\boldsymbol{y} = C\boldsymbol{a} + \boldsymbol{e} + \epsilon, \quad \boldsymbol{e} \in \mathrm{N}(\boldsymbol{0}, V_0), \quad \epsilon \in \mathrm{N}(\boldsymbol{0}, \sigma^2 I)$$

- Introduce parameters to describe the systematic effects,

$$\boldsymbol{e} = L_0\boldsymbol{d}, \quad V_0 = L_0 L_0^{\mathrm{T}}$$

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} C & L_0 \\ & I \end{bmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{d} \end{bmatrix} + \begin{bmatrix} \epsilon \\ \delta \end{bmatrix} \quad \epsilon \in \mathrm{N}(\boldsymbol{0}, \sigma^2 I), \quad \delta \in \mathrm{N}(\boldsymbol{0}, I)$$

Augmented system

$\tilde{\boldsymbol{y}} = \tilde{C}\tilde{\boldsymbol{a}} + \tilde{\epsilon}$, where

$$\tilde{\boldsymbol{y}} = \left[\begin{array}{c} \boldsymbol{y}/\sigma \\ \boldsymbol{0} \end{array} \right], \quad \tilde{C} = \left[\begin{array}{cc} C/\sigma & L_0/\sigma \\ & I \end{array} \right]$$

and

$$\tilde{\boldsymbol{a}} = \left[\begin{array}{c} \boldsymbol{a} \\ \boldsymbol{d} \end{array} \right], \quad \tilde{\epsilon} = \left[\begin{array}{c} \epsilon \\ \delta \end{array} \right] \quad \tilde{\epsilon} \in \mathrm{N}(\boldsymbol{0}, I)$$

- Eigenvalues

$$\hat{\tilde{\boldsymbol{y}}} = P\tilde{\boldsymbol{y}} = \left[\begin{array}{cc} P_{11} & P_{12} \\ P_{21} & P_{22} \end{array} \right] \left[\begin{array}{c} \boldsymbol{y}/\sigma \\ \boldsymbol{0} \end{array} \right]$$

$$\hat{\boldsymbol{y}} = P_{11}\boldsymbol{y}$$

- $n \leq \sum_j \lambda_j(P_{11}), \ \sum_j \lambda_j(P_{22}) \leq m$

Gaussian Processes

- Same extended model

$$\boldsymbol{y} = C\boldsymbol{a} + \boldsymbol{e} + \epsilon, \quad \boldsymbol{e} \in \mathrm{N}(\boldsymbol{0}, V_0), \quad \epsilon \in \mathrm{N}(\boldsymbol{0}, \sigma^2 I)$$

- $C_{ij} = b_j(t_i)$, $\mathrm{cov}(e, e') = k(t, t')$, e.g.

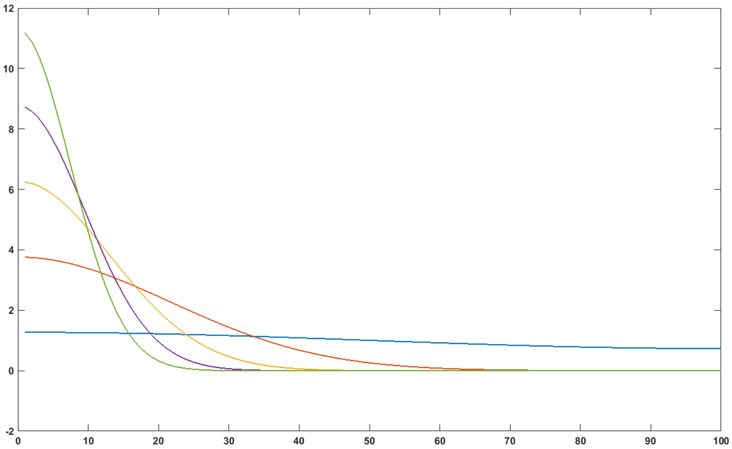
$$k(t, t') = \sigma_E^2 \exp\left\{-(t - t')^2/\tau^2\right\}$$

- Equally spaced t_i

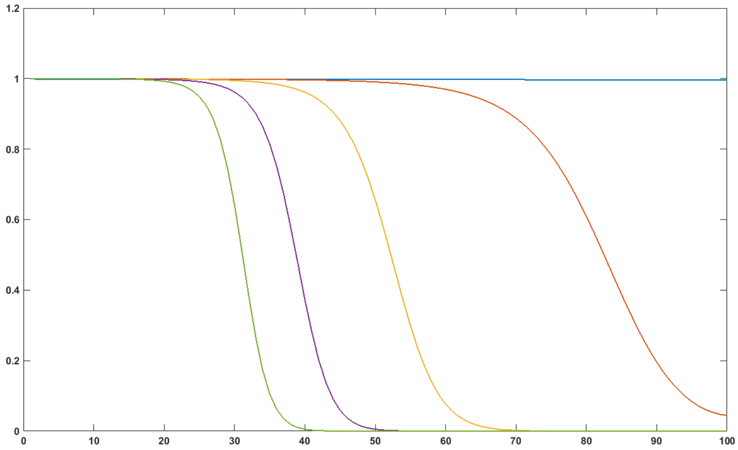
$$V = \sigma_E^2 \begin{bmatrix} 1 & v & v^4 & v^9 & v^{16} & \cdots \\ v & 1 & v & v^4 & v^9 & \cdots \\ v^4 & v & 1 & v & v^4 & \cdots \\ & & & \ddots & & \end{bmatrix}$$

NPL National Physical Laboratory

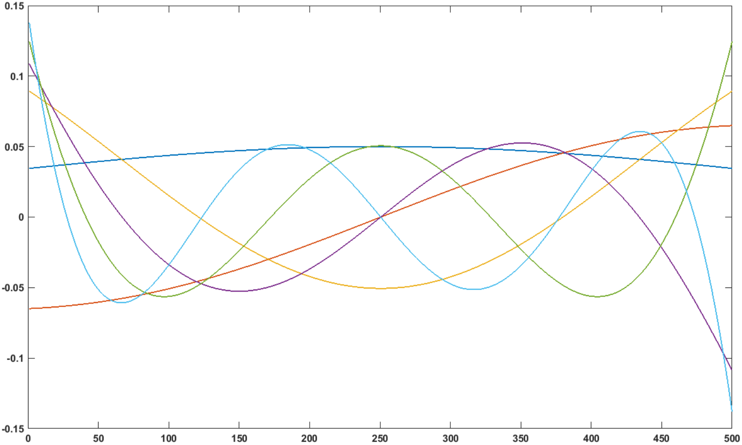
Eigenvalues of V for different τ



Eigenvalues of P_{11} for different τ



Eigenvectors of V



Eigenvectors as Chebyshev polynomials

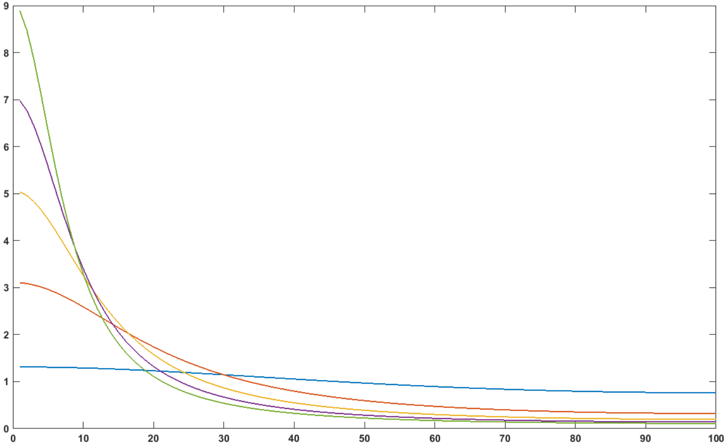
0.0838	−0.0002	0.0549	0.0009	0.0400	0.0018
0.0001	0.0724	−0.0004	−0.0485	−0.0013	−0.0366
−0.0077	0.0001	0.0697	0.0007	0.0461	0.0017
−0.0000	−0.0078	0.0001	−0.0687	−0.0009	−0.0449
0.0003	−0.0000	−0.0079	−0.0001	0.0681	0.0011
0.0000	0.0004	−0.0000	0.0080	0.0002	−0.0677
−0.0000	0.0000	0.0004	0.0000	−0.0081	−0.0002
−0.0000	−0.0000	0.0000	−0.0005	−0.0000	0.0081
0.0000	−0.0000	−0.0000	−0.0000	0.0005	0.0000
0.0000	0.0000	−0.0000	0.0000	0.0000	−0.0005

Chebyshev polynomials as eigenvectors

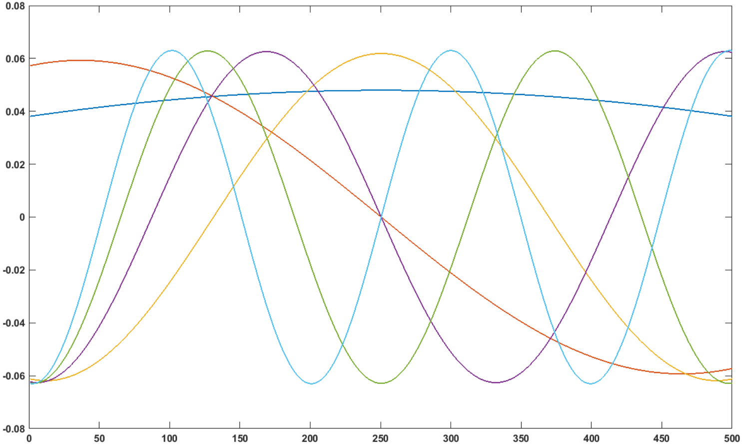
11.1174	0.0445	-8.8230	0.0275	-0.6102	0.0337
-0.0000	12.8329	0.1027	-9.1725	0.0586	-0.9364
1.1822	0.0047	12.3875	0.1628	-9.1967	0.0874
-0.0000	-1.4056	-0.0112	-12.5111	-0.2227	9.1655
0.0785	0.0003	1.4235	0.0180	12.6002	0.2820
-0.0000	-0.0869	-0.0007	-1.4731	-0.0250	-12.6561
0.0034	0.0000	0.0874	0.0011	1.5080	0.0320
0.0000	-0.0036	-0.0000	-0.0900	-0.0015	-1.5330
0.0001	0.0000	0.0036	0.0000	0.0920	0.0019
0.0000	-0.0001	-0.0000	-0.0037	-0.0001	-0.0935

NPL National Physical Laboratory

Eigenvalues of V, $k(t, t') \propto \exp\{-|t - t'|/\tau\}$



Eigenvectors of V, $k(t,t') \propto \exp\{-|t-t'|/\tau\}$



DIAL measurements and stack emissions

- DIAL: differential Absorption LIDAR
- Beams pointed at a plume emission
- Measures the cumulative absorption along the beam as a function of distance
- Absorption related to amount of pollutant along the beam
- Beam is stepped through a number of angles in a plane
- Goal: estimate the pollutant density of the plume

Air quality diagnostics

- Stacks at known locations
- Multi-species air quality sensors at known locations
- Prior profiles of species being emitted at different stacks
- Plume dispersion models
- Atmospheric chemistry models
- Met predictions: wind speed and direction
- Met data: wind speed and direction
- Goal: what is each stack is emitting as a function of time, alerts
- Goal: where to put air quality sensors (and which type) to provide best resolution
- Goal: determine air quality maps from the data and models
- Goal: find surrogate measurements, e.g., EO

Urban air quality diagnostics

- Prior profiles of emissions from different classes of vehicles, buildings
- Urban topography: maps, buildings, streets
- Environmental fluid dynamics
- Met data
- Traffic flow data: historical data, ANPR, speed cameras
- Multi-species air quality sensors at known locations
- Goal: determine posterior profiles of emission profiles
- Goal: predict air quality from traffic flow, met predictions

In-process measurement

- Workpiece ideal geometry at 20 degrees C specified, with tolerances
- Workpiece being manufactured: cutting, drilling, machining
- Measurements of the temperature at finite number of locations on the workpiece
- Measurements of the dimensions of a finite number of key features
- GOAL: use an FE model of artefact and the measurements to infer the workpiece shape at a stable 20 degrees
- Learn from an ensemble of workpieces
- Effective degrees of freedom associated with a FE model
- Minimise measurements required

Large engineering structures

- Aircraft wings, bridges
- FE model with many material parameters estimated
- Heterogeneous set of measurements: temperature, stress, strain, dimensions, tilt, accelerometers, windspeed
- Goal: use the FE model and data to improve estimates of the material parameters

Industry 4.0, digital twins

- Large scale models, simulations of factories
- Multiple streams of sensor data of actual behaviour
- Goal: assimilate data into models to improve predictability and decision-making

Guide to the Expression of Uncertainty in Measurement (GUM)

- Law of the propagation of uncertainty (1st and 2nd moments)

$$\boldsymbol{y} = C\boldsymbol{x}, \quad \mu_Y = C\mu_X, \quad V_Y = CV_XC^{\mathrm{T}}$$

- If $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}_X, V_X)$, then $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu}_Y, V_Y)$
- $\mathrm{N}(\boldsymbol{\mu}, V)$ is the maximum entropy distribution with mean $\boldsymbol{\mu}$ and variance V

Summarising a distribution, reconstructing an approximate distribution

- Given $p(x)$, calculate $S_k(p)$, $k = 1, \ldots, K$
- Given S_k, construct $p_0(x)$ such that $S_k(p_0) = S_k$
- For what class of distributions is $p_0 = p$
- S_k low order moments: mean, variance, skewness, kurtosis, etc.,
- S_k quantiles: 2.5, 5, 10, 50, 90, 95, 97.5
- $p(x) \rightarrow \mu_X, V_X \rightarrow \mathrm{N}(\mu_X, V_X)$

Maximum entropy distributions from moments

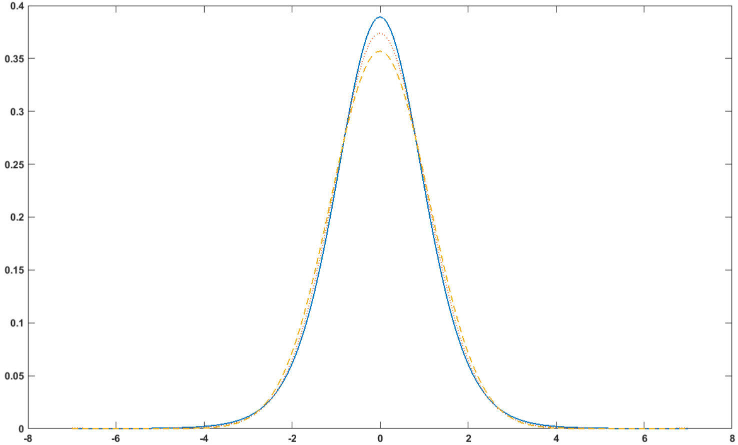
- Non-central moments

$$m_k = \int x^k p(x) \mathrm{d}x, \quad k = 0, \ldots, n$$

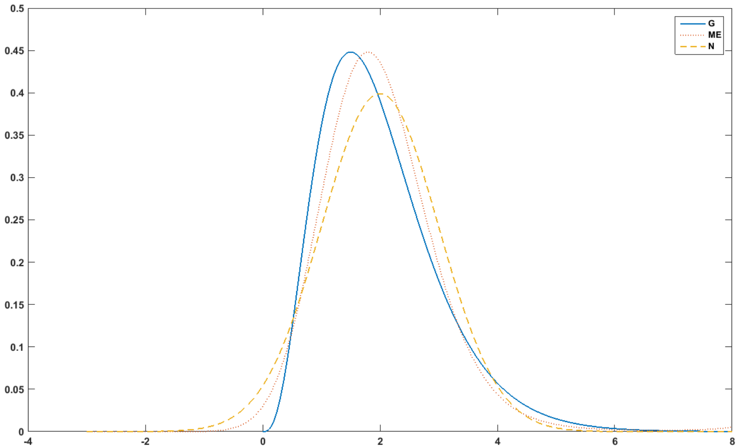
- Maximum entropy distribution satisfies

$$m_k = \int x^k \exp\left(\sum_{k=0}^{n} a_k x^k - 1\right) \mathrm{d}x$$

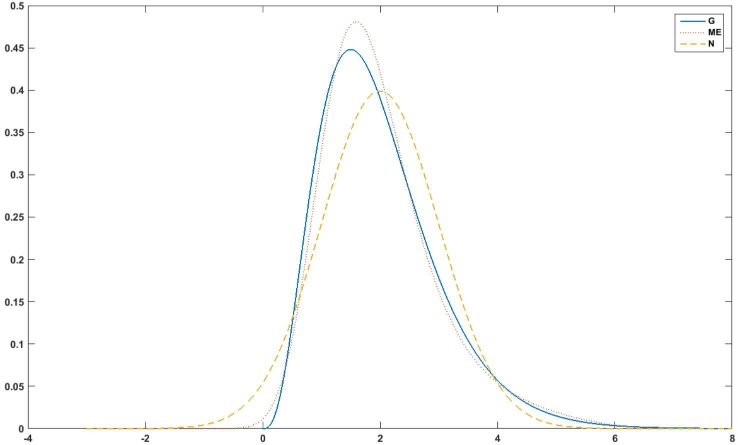
Reconstruction of a *t*-distribution



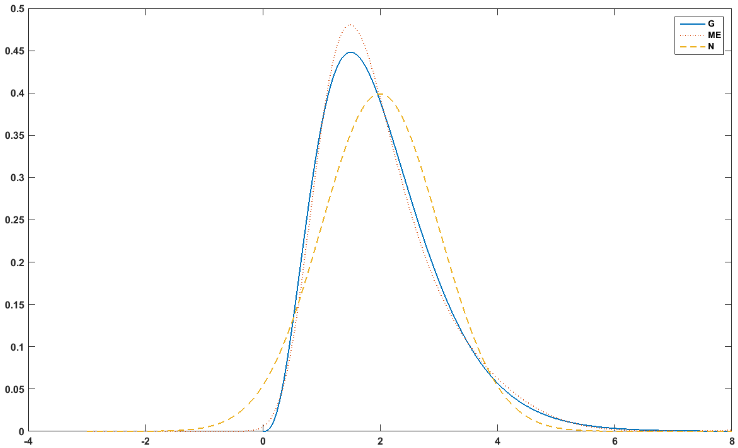
Reconstruction of a Gamma distribution, 3 moments



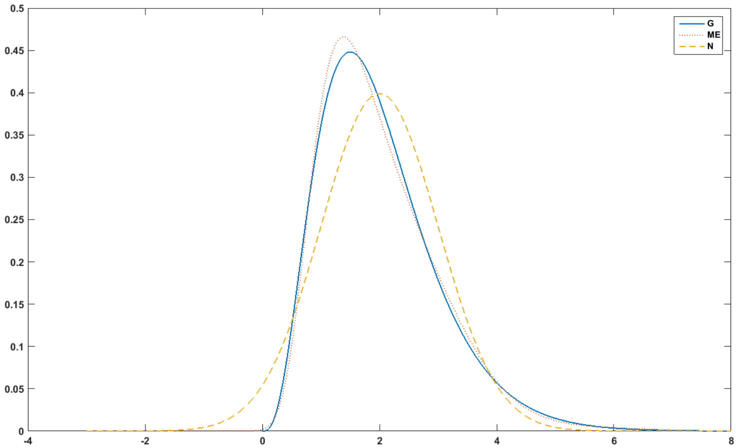
Reconstruction of a Gamma distribution, 4 moments



Reconstruction of a Gamma distribution, 5 moments



Reconstruction of a Gamma distribution, 6 moments



Maximum entropy distributions from quantile constraints

- Maximum entropy distribution given mean, variance and 2.5 and 97.5 quantiles
- Result is a discontinuous piecewise Gaussian

A more general problem

Given a space of probability P distributions with a prior on P, choose quantiles Q_k and reconstruction scheme R to minimise the expected value of

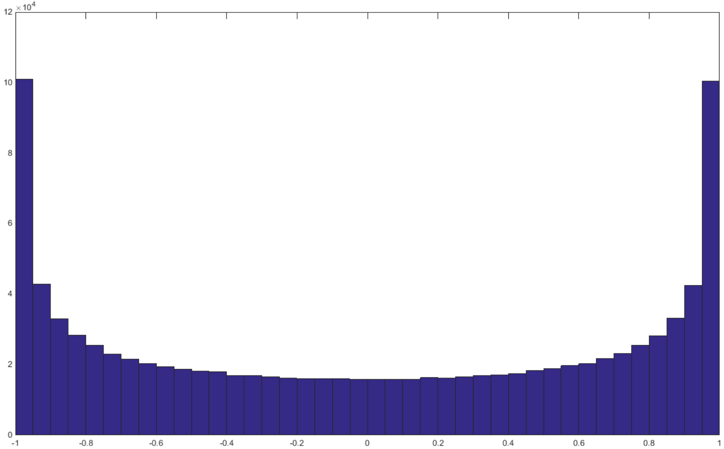
$$D(p||R(Q_k(p))$$

(or some other measure).

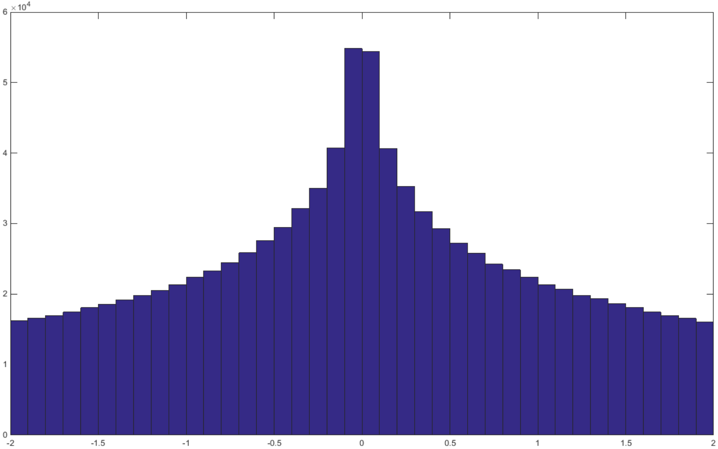
Chebyshev-type inequalities

- Suppose $y = \sum_1^n x_j$ where x_j has mean and variance μ_j, σ_j^2, derive tight estimates of the quantiles associated with y.
- What can be said if we know more: higher moments, symmetry, unimodality

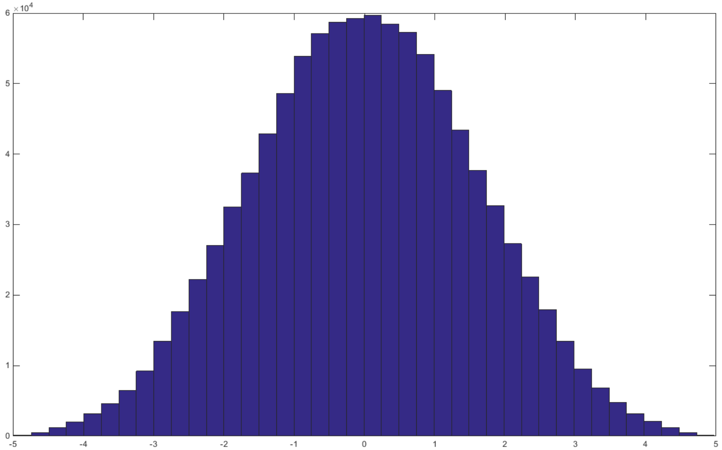
Arcsine distribution



Sum of 2 arcsine variates



Sum of 5 arcsine variates



Other statistical interests

- Approximate Bayesian computation
- Linear Bayes
- Imprecise probability
- Probabilistic numerics