Computational tools II: Markov Chain Monte Carlo (MCMC) and the Gibbs sampler

Goal: Represent probability distributions by random samples.

Hence, we have to be able to generate (usually dependent!) samples from a given distribution $p(x)$. In the application to Bayesian models case x is set of parameters and p the posterior.

Basic method: Transformation method and rejection method with proposal density

- Problem: Need random variables with density $p(x)$ (target density), have random variables with density $q(x)$ (proposal density).

- **Transformation method:**

  Find a transformation $x = f(y)$ such that the distribution of x is $p(x)$.

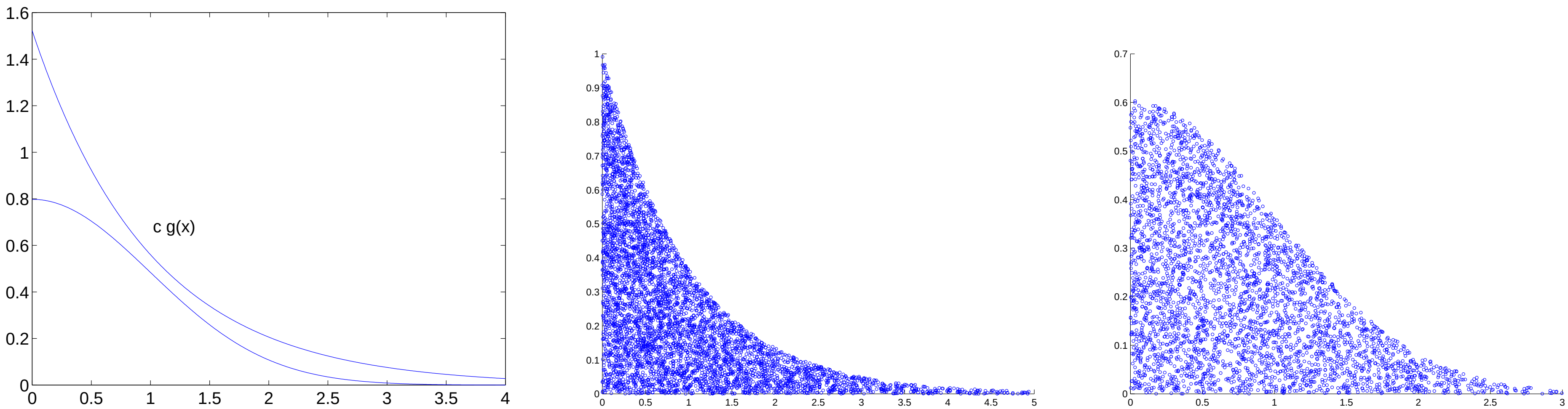
  Let $F(z) = P(x \leq z)$ with density $p(x) = F'(x)$. Let $y \sim U(0, 1)$ a random variable with uniform density. Then the transformed $x = F^{-1}(y)$ has density $p(x)$.

- **Rejection method:**

  Assume $\frac{p(x)}{q(x)} \leq c$. Generate two independent random variables $x \sim q(x)$ and $u \sim U(0, 1)$. If $u \leq \frac{p(x)}{cq(x)}$ accept x. Otherwise start again.

Example: Exponential $\to$ Normal

- We can get *positive normal (Gaussian)* random variables with density $p(x) = \frac{2}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ for $0 \le x < \infty$ by the *rejection method* using exponentially distributed. A good candidate is $c = \sqrt{2e/\pi}$ and $\frac{p(x)}{cq(x)} = \exp(-(x-1)^2/2)$.



Note: The rejection method can also be applied to the case where we know the desired distribution only up to a normalisation constant, i.e. $p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$ with unknown Z.

Markov Chain Monte Carlo

- It easy to sample from simple low dimensional distributions by the transformation or the rejection methods. But this doesn't work well for higher dimensions.

- General Strategy: Construct a Markov chain with a transition probability $T(y|x)$ that has $p(x)$ as its stationary distribution.

- Let us assume that there is only a single stationary distribution and that any initial distribution converges to it. Then, asymptotically (that is if we wait long enough), the distribution of samples X_t drawn from the Markov chain is very close to $p(x)$.

Stationary distributions

Let $p_t(x)$ denote the marginal distribution of X_t. The update of the marginal distribution given by

$$p_{t+1}(x) = \int T(x|y) p_t(y) \, dy$$

The *stationary distribution* must fulfil stationarity

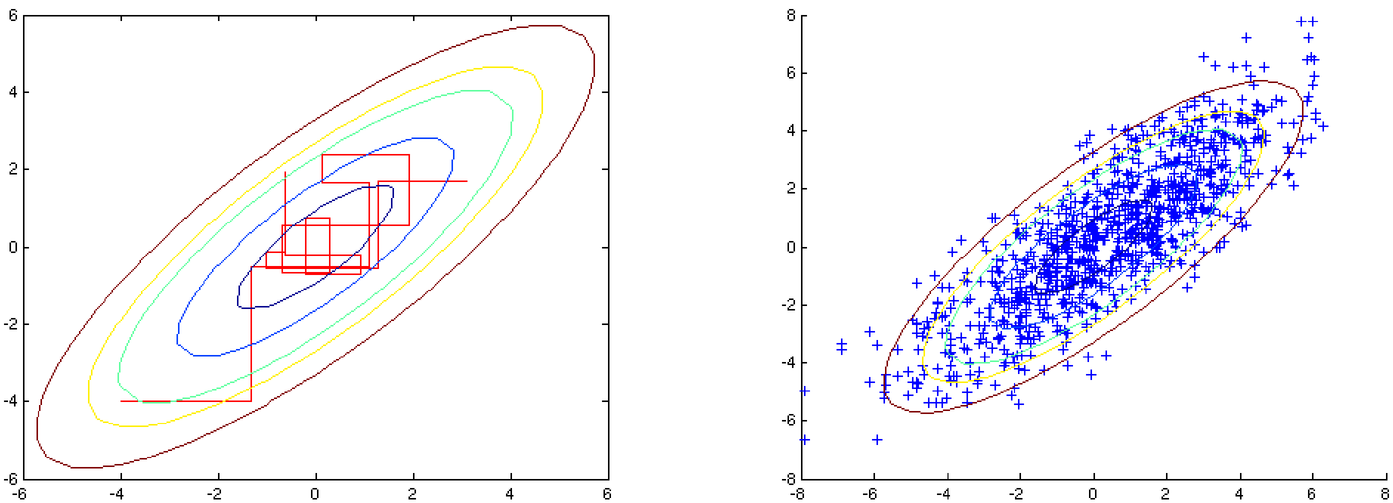
$$p(x) = \int T(x|y) p(y) \, dy$$

Hence, we should find transition probabilities which leave our target distribution invariant.

Gibbs sampling

is easily applied when one can sample from the conditional probabilities $p(x_i|\mathbf{x}_{-i})$ where $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$. At step $\tau + 1$, one cycles through the components of $\mathbf{x}$ and samples

$$
\begin{array}{rcl}
x_1^{\tau+1} & \sim & p(x_1|x_2^\tau, x_3^\tau, \ldots, x_N^\tau) \\
x_2^{\tau+1} & \sim & p(x_2|x_1^{\tau+1}, x_3^\tau, \ldots, x_N^\tau) \\
\ldots & \ldots & \ldots \\
x_j^{\tau+1} & \sim & p(x_j|x_1^{\tau+1}, \ldots, x_{j-1}^{\tau+1}, x_{j+1}^\tau, \ldots, x_N^\tau) \\
\ldots & \ldots & \ldots \\
x_N^{\tau+1} & \sim & p(x_N|x_1^{\tau+1}, \ldots, x_{N-1}^{\tau+1})
\end{array}
$$



Application: Change point model

Disasters can occur at years $i \in \{1, 2, \ldots, n\}$. Number of disasters are distributed as a Poisson variable, ie $p(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$. But the rate of disasters change from λ_1 to λ_2 at unknown **change point** $K \in \{1, 2, \ldots, n\}$.

To estimate K we assume the following hierarchical Bayesian model

- K has a discrete prior distribution $p(K)$.

- Given K and $\lambda_{1,2}$, the data are independent $x_i \sim e^{-\lambda}\frac{\lambda^x}{x!}$.

- The rates $\lambda_{1,2}$ are independent with $\lambda_{1,2} \sim \text{Gamma}(a_{1,2}, \eta_{1,2})$ density. $\eta_{1,2}$ are hyperparameters and $a_{1,2}$ are known.

- $\eta_{1,2}$ are independent hyperparameters $\eta_{1,2} \sim$ Gamma$(b_{1,2}, c_{1,2})$ with known $b_{1,2}$ and $c_{1,2}$.

Note that the Gamma density is given by

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

with $E[X] = \frac{\alpha}{\beta}$ and $Var[X] = \frac{\alpha}{\beta^2}$.

Problem: Given a set of observations $\mathbf{x} = (x_1, \ldots, x_n)$ over n years, draw samples from the **posterior distribution** $p(K, \eta, \lambda|\mathbf{x})$.

- Joint distribution

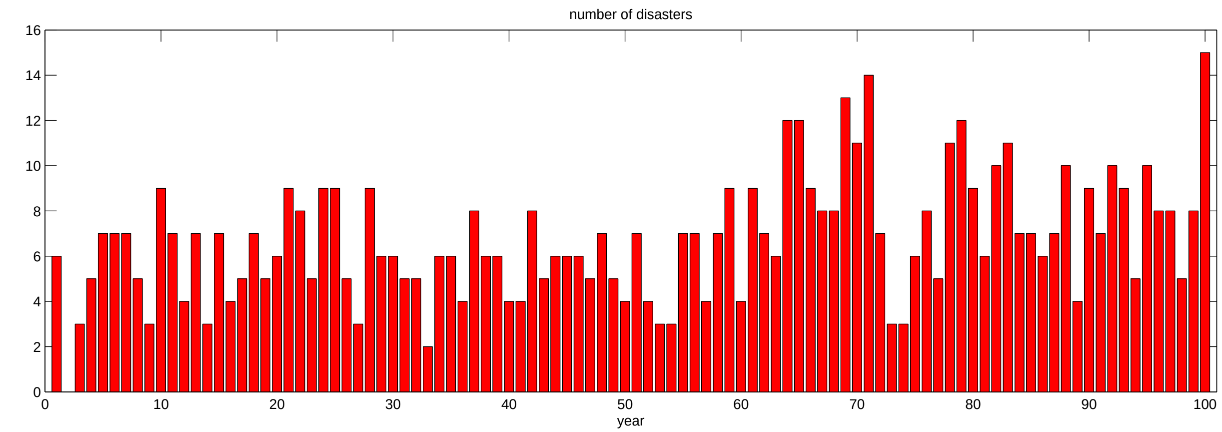
$$p(\mathbf{x}, \lambda_{1,2}, \eta_{1,2}, K) = p(\mathbf{x}|\lambda_{1,2}, K)p(\lambda_{1,2}|\eta_{1,2})p(\eta_{1,2})p(K) =$$

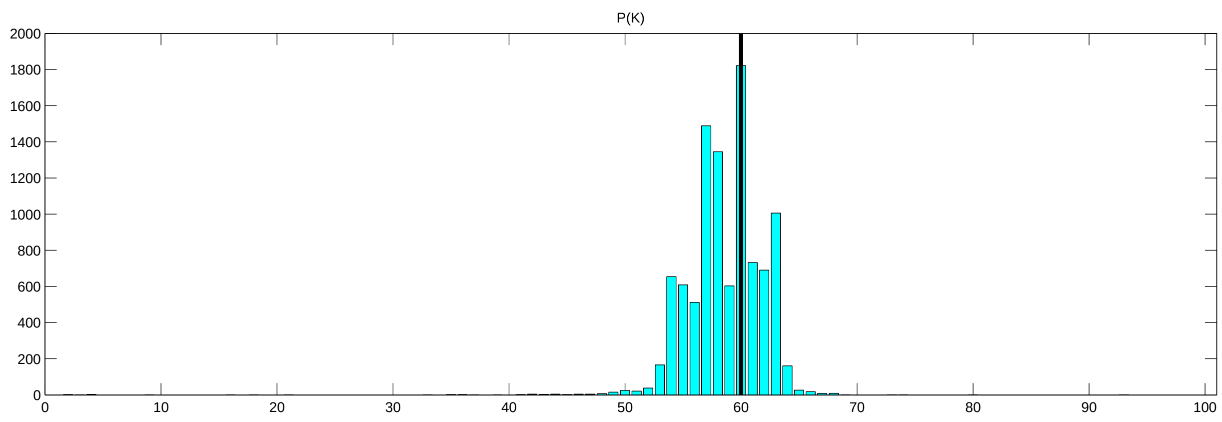
$$\prod_{i=1}^{K} e^{-\lambda_1}\frac{\lambda_1^{x_i}}{x_i!} \times \prod_{K+1}^{n} e^{-\lambda_2}\frac{\lambda_2^{x_i}}{x_i!} \times$$
$$\times\frac{\eta_1^{a_1}}{\Gamma(a_1)}\lambda_1^{a_1-1}e^{-\eta_1\lambda_1} \times \frac{\eta_2^{a_2}}{\Gamma(a_2)}\lambda_2^{a_2-1}e^{-\eta_2\lambda_2} \times$$
$$\times\frac{c_1^{b_1}}{\Gamma(b_1)}\eta_1^{b_1-1}e^{-c_1\eta_1} \times \frac{c_2^{b_2}}{\Gamma(b_2)}\eta_2^{b_2-1}e^{-c_2\eta_2} \times$$
$$\times p(K)$$

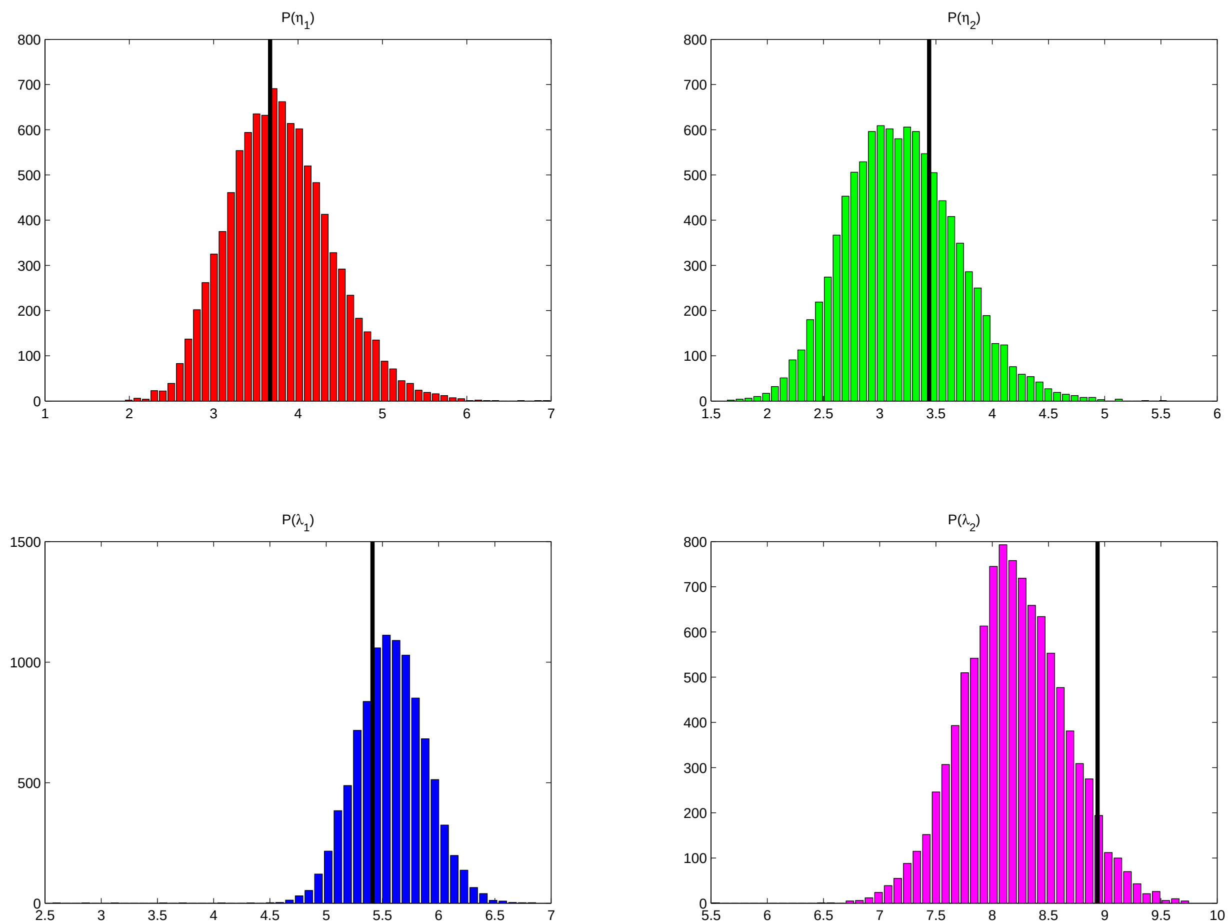
- Conditional distributions for Gibbs sampler

$$\begin{aligned}
\lambda_2|\lambda_1, \eta_{1,2}, K, \mathbf{x} &\sim \text{Gamma}(a_2 + \sum_{K+1}^{n} x_i, n - K + \eta_2) \\
\eta_1|\lambda_{1,2}, \eta_2, K, \mathbf{x} &\sim \text{Gamma}(a_1 + b_1, \lambda_1 + c_1) \\
K|\lambda_{1,2}, \eta_{1,2}, \mathbf{x} &\sim \text{const} \times p(K)e^{-K(\lambda_1-\lambda_2)}(\lambda_1/\lambda_2)^{\sum_{i=1}^{K} x_i}
\end{aligned}$$

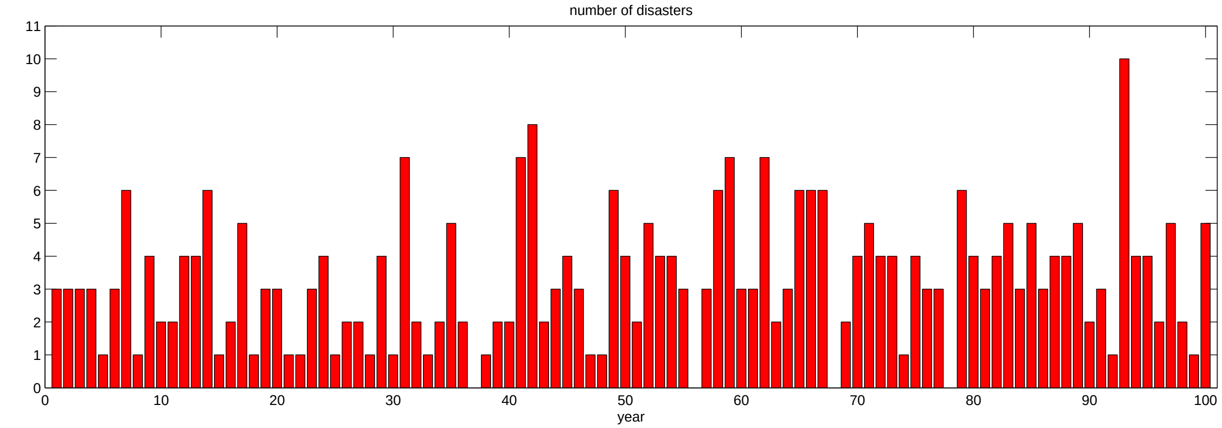
Simulations

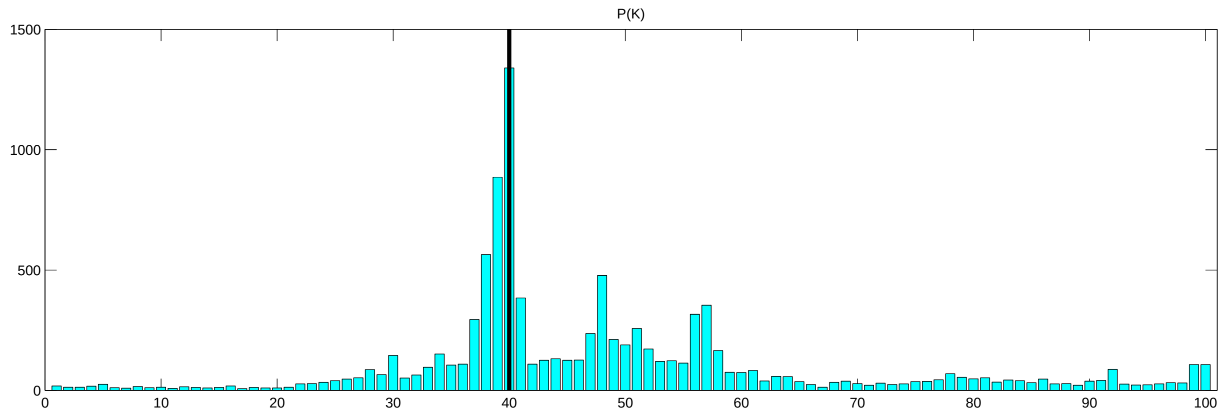


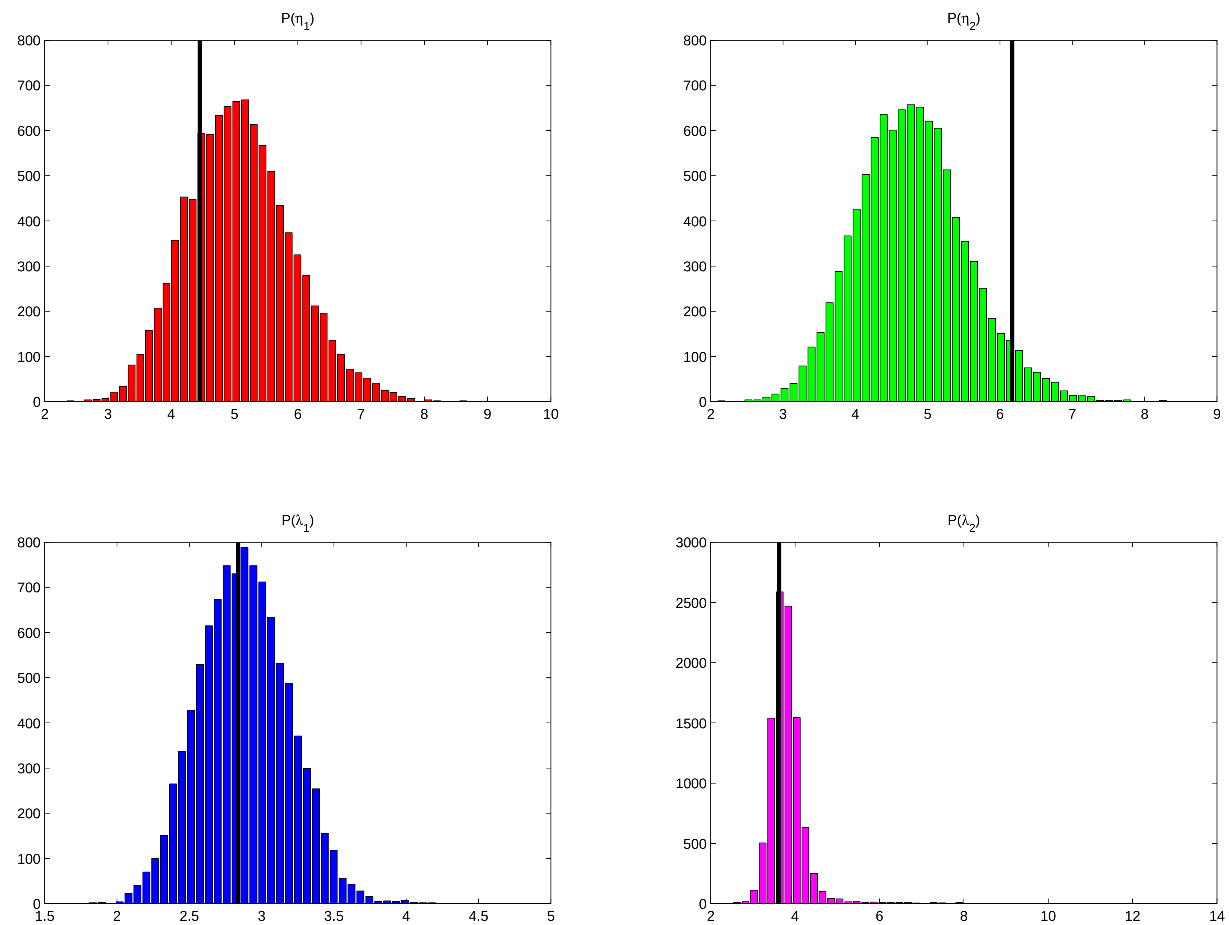



P(η1)
P(η2)
P(λ1)
P(λ2)

with somewhat more similar λ_{12}







P(η1)
P(η2)
P(λ1)
P(λ2)

Factor analysis

Observed data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are explained by a set of latent variables $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$. The model is in matrix notation

$$\mathbf{X} = \mathbf{M} + \mathbf{\Lambda}\mathbf{F} + \mathbf{E}$$

- $d =$ dimensionality of data. $n =$ number of observations.
- The data matrix is $d \times n$, the *factor loadings* matrix $\mathbf{\Lambda}$ is $d \times q$, the factors $\mathbf{F}$ are $q \times n$ and the error matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ is $d \times n$.
- The noise is $E[\mathbf{E}\mathbf{E}^\top] = \mathbf{\Psi} = \text{diag}\,(\psi_1^2, \ldots, \psi_d^2)$
- $p(\mathbf{X}|\mathbf{F}, \mathbf{\Lambda}, \mathbf{M}, \mathbf{\Psi}) = \mathcal{N}(\mathbf{X}|\mathbf{M} + \mathbf{\Lambda}\mathbf{F}, \mathbf{\Psi})$
- $p(\mathbf{f}_i) = \mathcal{N}(0, \mathbf{\Sigma}_f)$. Often $\mathbf{\Sigma}_f = \mathbf{I}$ is chosen.
- Total likelihood of data $p(\mathbf{X}|\mathbf{\Lambda}, \mathbf{M}, \mathbf{\Psi}) = \mathcal{N}(\mathbf{X}|\mathbf{M}, \mathbf{\Lambda}\mathbf{\Sigma}_f\mathbf{\Lambda}^\top + \mathbf{\Psi})$

Non - Bayesian Inference

- One can use the EM algorithm to estimate Maximum Likelihood estimators of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$.

- *Sparsity* of factor loadings: Use nonidentifiability and apply *rotations* with orthogonal $\mathbf{Q}$ to trained loading matrix $\mathbf{\Lambda}$: $\mathbf{\Lambda}_{rot} = \mathbf{\Lambda Q}$ to create sparse $\mathbf{\Lambda}_{rot}$.

  Use sparsity penalty. e.g.

$$\sum_{k=1}^{q} \sum_{l=1}^{d} \tanh\left(\alpha \lambda_{lk}^2\right)$$

  or *procrustes* rotation with penalty

$$\sum_{k=1}^{q} \sum_{l=1}^{d} (\lambda_{lk} - \tau_{lk})^2$$

  where τ_{lk} is a *target* matrix.

Bayesian inference (E. Fokoue)

- *Bayesian* approach: Introduce sparsity prior, e.g. by products of student densities

$$p(\lambda_{lk}|\alpha, \beta) \propto \frac{1}{\left(\beta + \frac{1}{2}\lambda_{lk}^2\right)^{\alpha+\frac{1}{2}}}$$

which has high probability densities at the coordinate axes:

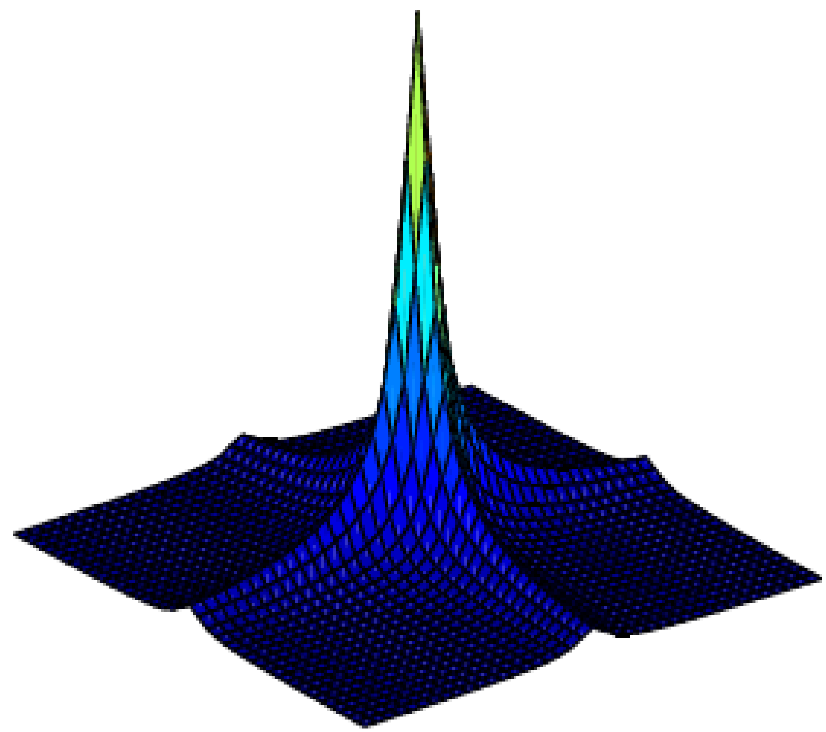


Figure 1: The 2-dimensional marginal prior for a row Λ_i

Let $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Psi})$: Sampling from $p(\boldsymbol{\theta}|\mathbf{X})$ is not feasible:

Posterior has complicated dependency on $\boldsymbol{\Lambda}$

$$p(\Lambda|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) = \mathcal{N}(\mathbf{X}|\mathrm{M}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})p(\boldsymbol{\Lambda}) \propto$$
$$|\boldsymbol{\Lambda}\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{X} - \mathrm{M})^\top(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})^{-1}(\mathbf{X} - \mathrm{M})\right]$$

Data Augmentation

Introducing the auxiliary variables δ_{lk} with

$$
\begin{aligned}
p(\lambda_{lk}|\delta_{lk}) &= \mathcal{N}(0, 1/\delta_{lk}) \\
p(\delta_{lk}|\alpha, \beta) &= \frac{\delta_{lk}^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\delta_{lk}}
\end{aligned}
$$

The marginal distribution is just

$$
p(\lambda_{lk}|\alpha, \beta) \propto \frac{1}{\left(\beta + \frac{1}{2}\lambda_{lk}^2\right)^{\alpha+\frac{1}{2}}}
$$

- Try to sample from $p(\boldsymbol{\Delta}, \boldsymbol{\theta}, \mathbf{F}|\mathbf{X})$ instead.

- Gibbs sampler: Alternate sampling between $p(\mathbf{F}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\Delta})$, $p(\boldsymbol{\theta}|\boldsymbol{\Delta}, \mathbf{F}, \mathbf{X})$ and $p(\boldsymbol{\Delta}|\boldsymbol{\theta}, \mathbf{F}, \mathbf{X})$

- Conditional of factors is a Gaussian

$$\mathbf{f}_i | \mathbf{x}_i, \mathbf{\Lambda}, \mathbf{\Psi} \sim \mathcal{N}\left((\mathbf{I}_q + \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1} \mathbf{\Lambda})^{-1} \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1} \mathbf{x}_i, (\mathbf{I}_q + \mathbf{\Lambda}^\top \mathbf{\Psi}^{-1} \mathbf{\Lambda})^{-1} \right)$$

- Conditional of $\mathbf{\Lambda}$

$$
\begin{aligned}
p(\mathbf{\Lambda}|\mathbf{X}, \mathbf{F}, \mathbf{M}, \mathbf{\Delta}) \propto p(\mathbf{X}|\mathbf{F}, \mathbf{M}, \mathbf{\Lambda}, \mathbf{\Delta}) p(\mathbf{\Lambda}|\mathbf{\Delta}) = \\
\mathcal{N}(\mathbf{X}|\mathbf{M} + \mathbf{\Lambda}\mathbf{F}, \mathbf{\Psi}) p(\mathbf{\Lambda}|\mathbf{\Delta}) \propto \\
\exp\left[-\frac{1}{2}(\mathbf{X} - (\mathbf{M} + \mathbf{\Lambda}\mathbf{F}))^\top \mathbf{\Psi}^{-1} (\mathbf{X} - (\mathbf{M} + \mathbf{\Lambda}\mathbf{F}))\right] p(\mathbf{\Lambda}|\mathbf{\Delta})
\end{aligned}
$$

  is also Gaussian !

- Finally

$$
p(\mathbf{\Delta}|\mathbf{\Lambda}, \mathbf{X}, \mathbf{F}, \mathbf{M})
$$

  is a product of *Gamma* distributions.

A model for collaborative filtering

(U Paquet, B Thomson, O Winther; *A hierarchical model for ordinal matrix factorization*, Statistics and Computing)

- r_{mn} = Rating of customer n on item (e.g. movie) m. We have $r_{mn} \in 1, \ldots, R$
- Introduce ideal latent variable f with $p(r|f) = 1$ if $b_r \leq f \leq b_{r+1}$, where $-\infty = b_1 < b_2 < \ldots < b_{R+1} = \infty$ and $p(r|f) = 0$, else.
- The latent variable f becomes noisy using $p(f|h) = \mathcal{N}(f; h, 1)$. This leads to

$$p(r_{mn}|h_{mn}) = \prod_r \Big[\Phi(h_{mn} - b_r) - \Phi(h_{mn} - b_{r+1})\Big]^{1_{r_{mn}=r}}$$

and the total likelihood is

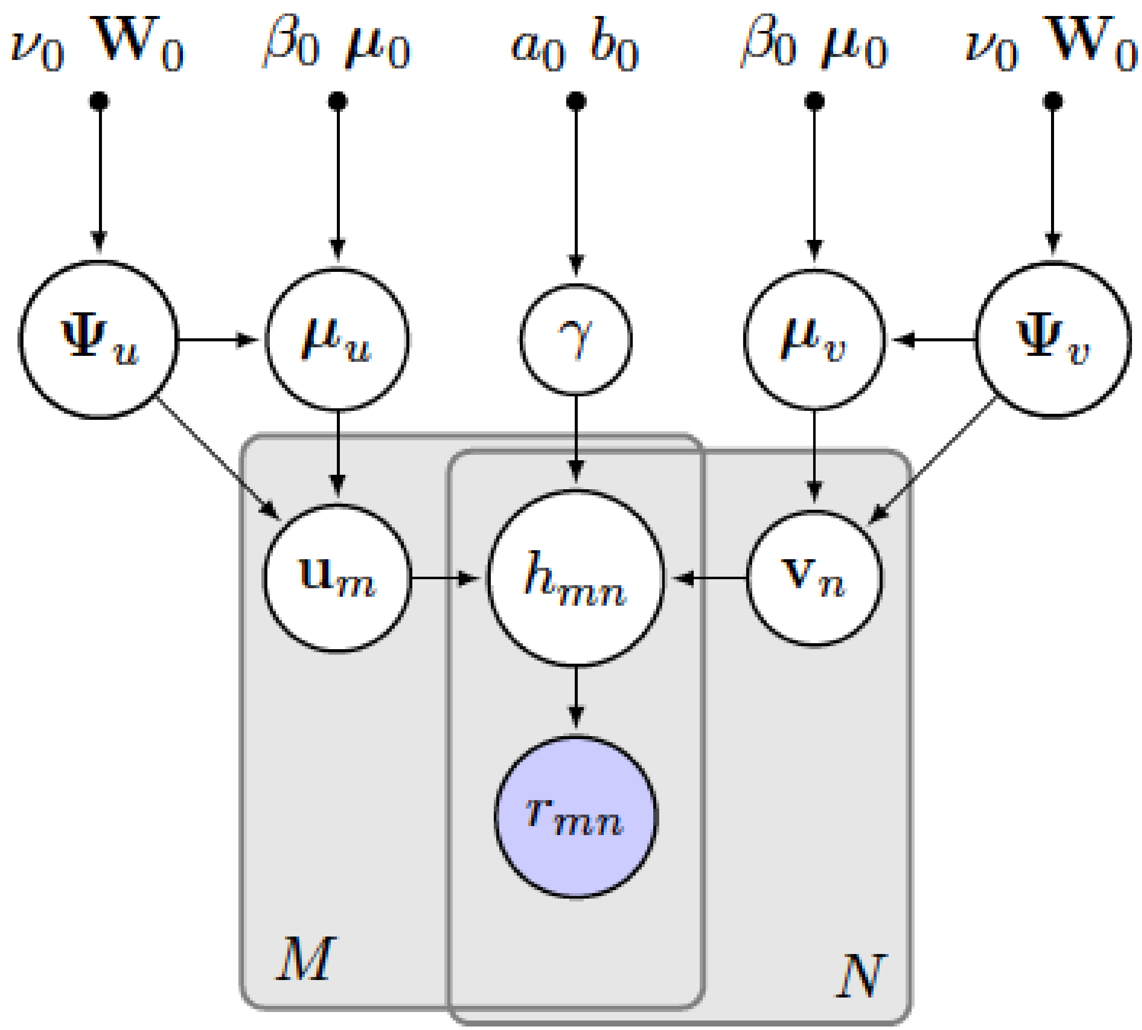
$$p(D|H) = \prod_{m,n} p(r_{mn}|h_{mn})$$

- **Low rank matrix factorization:** $h_{mn} = \mathbf{u}_m^\top \mathbf{v}_n + \epsilon_{mn}$ with $\epsilon_{mn} \sim \mathcal{N}(0, \gamma^{-1}$ i.i.d. Gaussian noise.

- $\mathbf{u}_m$ and $\mathbf{v}_n$ are factors of length K (small) corresponding to item m and customer n.

- Priors $p(\mathbf{u}_m|\boldsymbol{\mu}_u, \boldsymbol{\Psi}_u) = \mathcal{N}(\mathbf{u}_m; \boldsymbol{\mu}_u, \boldsymbol{\Psi}_u)$ and $p(\mathbf{v}_n|\boldsymbol{\mu}_v, \boldsymbol{\Psi}_v) = \mathcal{N}(\mathbf{v}_n; \boldsymbol{\mu}_v, \boldsymbol{\Psi}_v)$

- $p(\boldsymbol{\mu}_{u,v}, \boldsymbol{\Psi}_{u,v}) =$ Normal–Wishart priors. $p(\gamma)$ is a Gamma prior.



ν_0 $\mathbf{W}_0$
β_0 $\boldsymbol{\mu}_0$
a_0 b_0
β_0 $\boldsymbol{\mu}_0$
ν_0 $\mathbf{W}_0$
$\boldsymbol{\Psi}_u$
$\boldsymbol{\mu}_u$
γ
$\boldsymbol{\mu}_v$
$\boldsymbol{\Psi}_v$
$\mathbf{u}_m$
h_{mn}
$\mathbf{v}_n$
r_{mn}
M
N

- **Gibbs sampler:** We get e.g.

$$\mathbf{u}_m \sim \mathcal{N}\left(\mathbf{u}_m; \mathbf{\Sigma}_m \left[\Psi_u \boldsymbol{\mu}_u + \gamma \sum_{n \in \Omega(m)} h_{mn} \mathbf{v}_n\right], \mathbf{\Sigma}_m\right)$$

with

$$\mathbf{\Sigma}_m = \left(\boldsymbol{\psi}_u + \gamma \sum_{n \in \Omega(m)} \mathbf{v}_n \mathbf{v}_n^\top\right)^{-1}$$

Setting $\mu = \mathbf{u}^T \mathbf{v}$, we also have

$$p(r|f)p(f|h)p(h|\mu,\gamma) = \left[\Theta(b_{r+1} - f) - \Theta(b_r - f)\right] \mathcal{N}(f; h, 1)\mathcal{N}(h; \mu, \gamma^{-1})$$

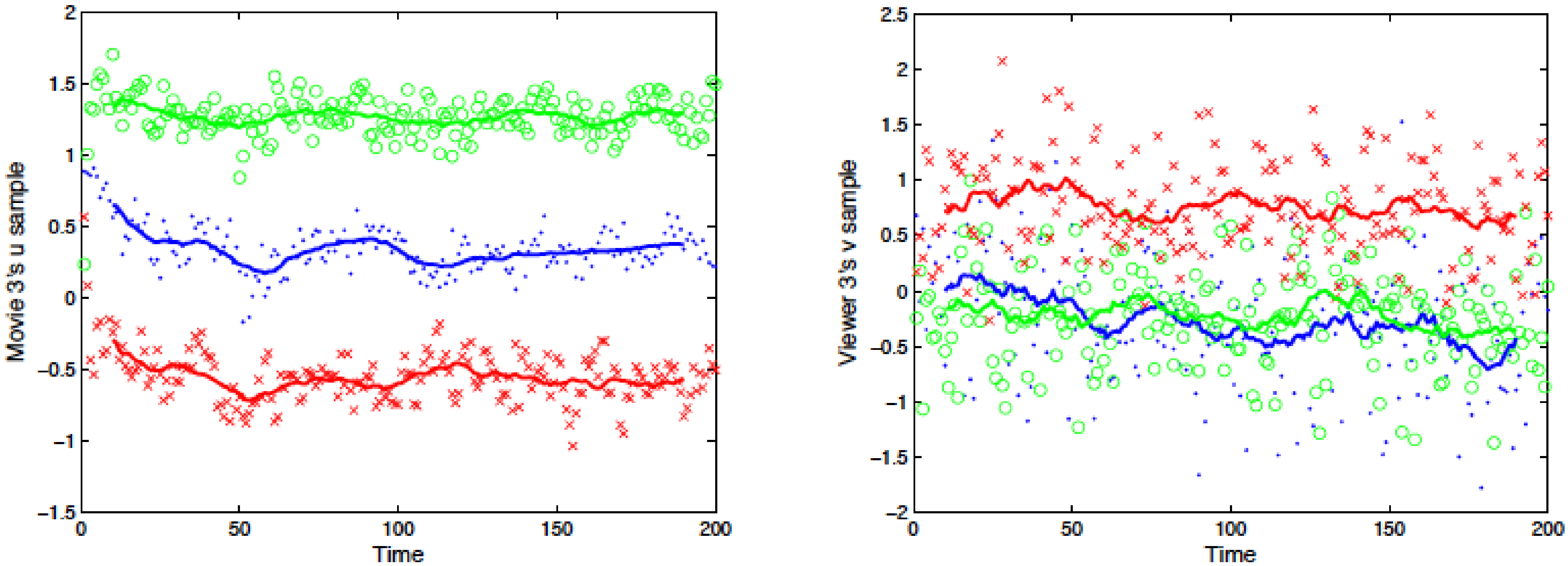


Figure 6: The samples for six of the five million model parameters required for a "small" model with $K = 10$. The samples for the first three components of $\mathbf{u}_3$ for movie 3, i.e. u_{13}, u_{23}, and u_{33}, are shown at the top. Movie 3 had $\Omega(3) = 2011$ ratings. The samples for the first three components of $\mathbf{v}_3$ are shown at the bottom.Viewer 3 rated $\Pi(3) = 97$ movies. The overlaid lines indicate a windowed average over 20 samples.

- **Application:** *Netflix* data set with $N = 480,189$ users and $M = 17,770$ movies and 100 Million ratings. Test on hold–out data with 3 Million user–movie pairs gave a root mean square error of $RMS = 0.8913$ compared to the original algorithm of Netflix which gave $RMS = 0.9514$. The optimum (award winning algorithm) based on another method had $RMS = 0.8567$.