Inter-Disciplinary Project

Estimation of Probability Density Functions and Graphical Models Using Regularized Sparse Grids

Karthikeya Sampa Subbarao

28 October 2015

Agenda

- Density Estimation
- Sparse Grids with Modified Linear Basis
- Density Estimation using Sparse Grids
- Performance Improvement Strategies
    - ✓ Dynamic Learning Rate
    - ✓ Grid Coarsening
- Experiments and Results

Density Estimation

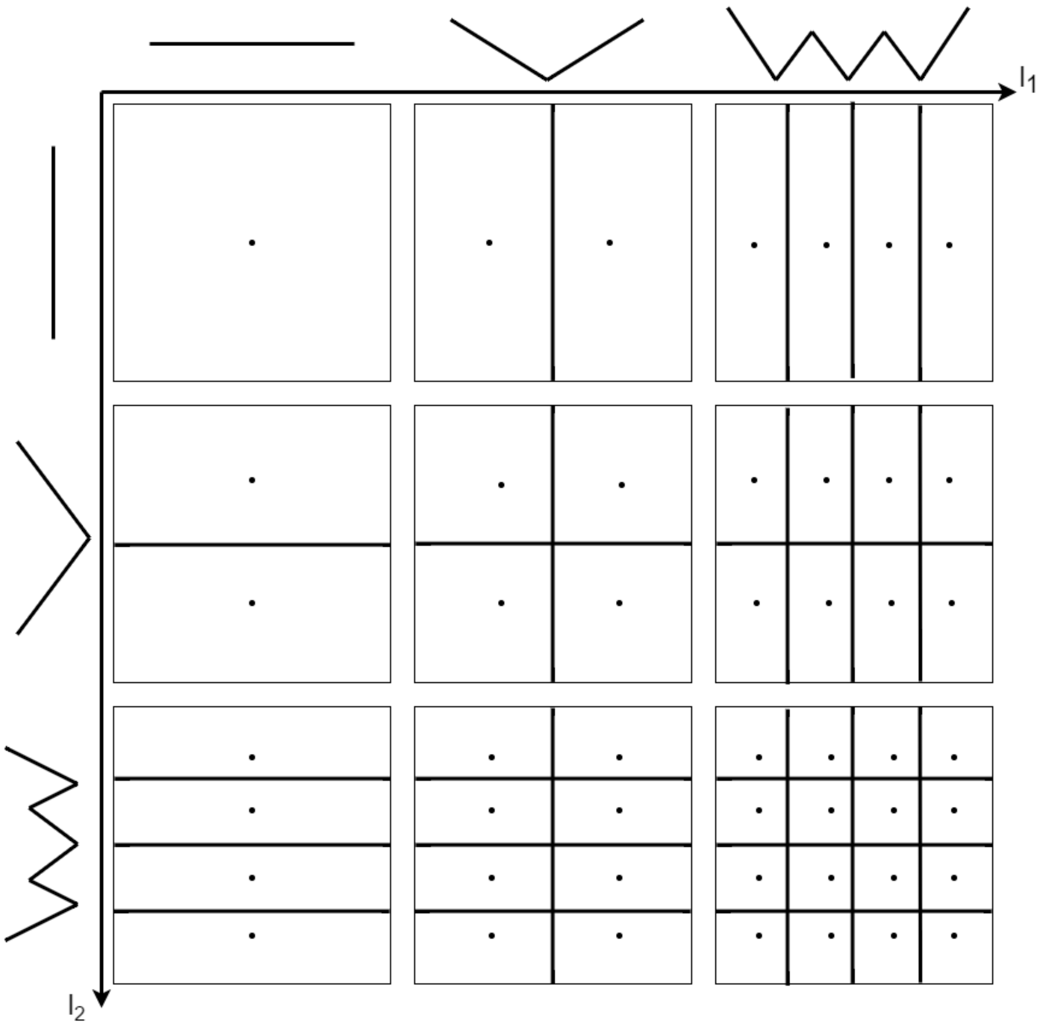
$$D = \{x_1, x_2, \dots, x_M\}, \qquad where\ x_i \in \mathbb{R}^d$$
$$drawn\ from\ a\ distribution\ with\ probability\ density\ function\ p(X).$$

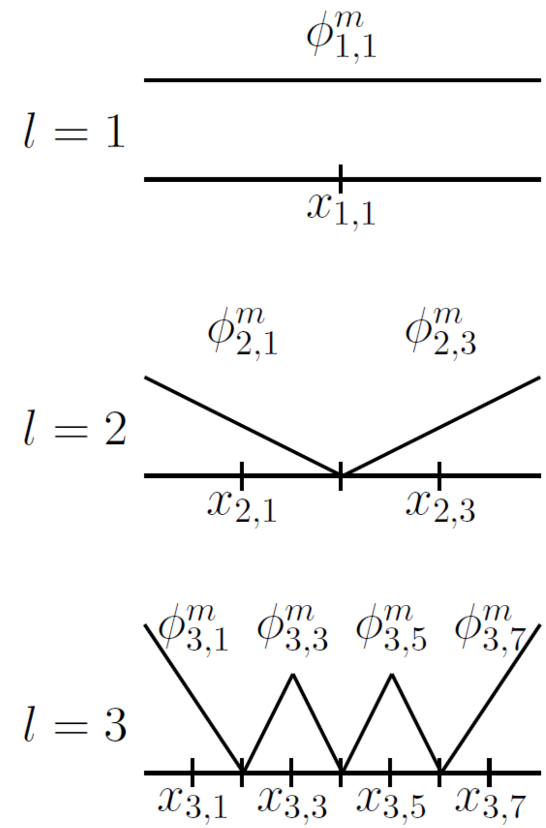
Estimate $\hat{p}$ of the probability density function p based on the dataset D.

- ❑ Parametric Density Estimation
- ❑ Non-parametric Density Estimation
  - Kernel Density Estimation (KDE)

Data centric density estimation faces curse of dimensionality

Sparse Grids with Modified Linear Basis





Source: Benjamin Peherstorfer. Model Order Reduction of Parametrized Systems with Sparse Grid Learning Techniques. October 2013.

Analysis of Variance (ANOVA)

Decomposition of d-dimensional function h

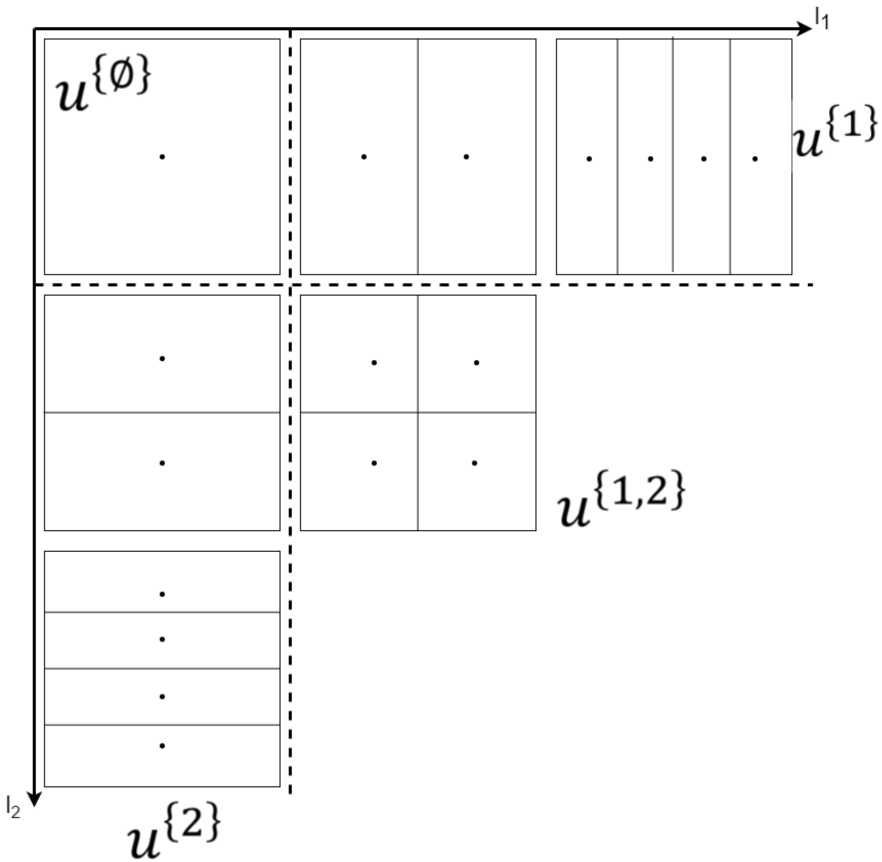
$$h(x_1, \ldots, x_d) = h_0 + \sum_{j_1}^{d} h_{j_1}(x_{j_1}) + \sum_{j_1<j_2}^{d} h_{j_1,j_2}(x_{j_1}, x_{j_2})$$

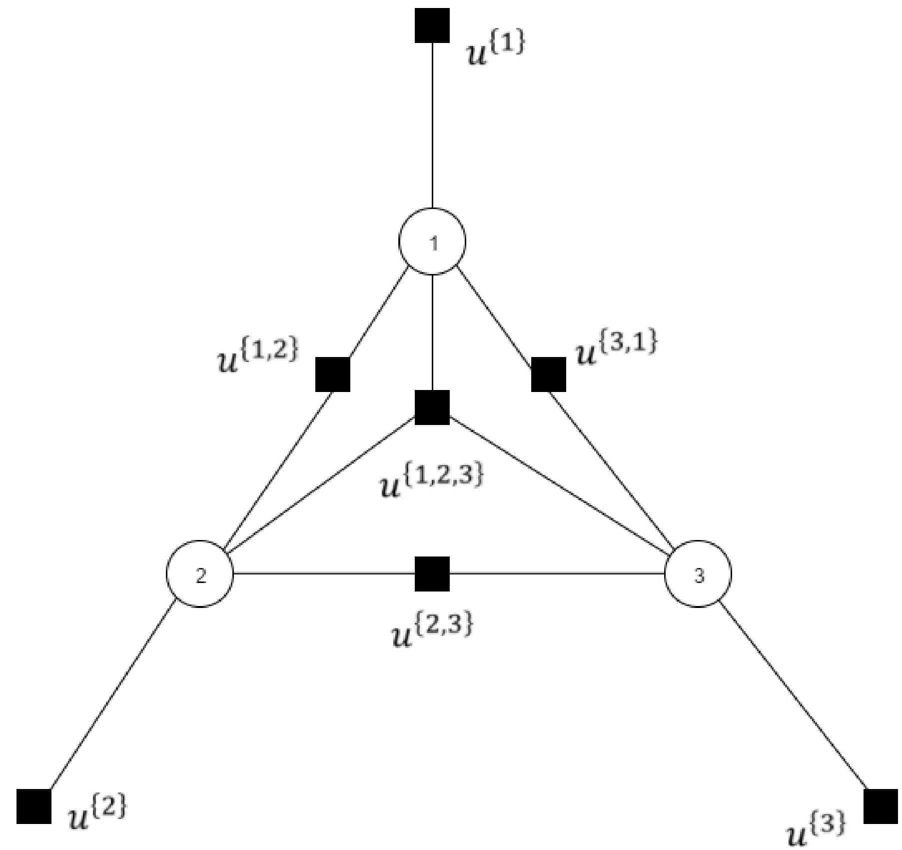
$$+ \sum_{j_1<j_2<j_3}^{d} h_{j_1,j_2,j_3}(x_{j_1}, x_{j_2}, x_{j_3}) + \cdots + h_{j_1,\ldots,j_d}(x_{j_1}, \ldots, x_{j_d}))$$

Decomposition of sparse grid function space with respect to hierarchical subspace

$$v_l = u^{\{\emptyset\}} \bigoplus_{i=1}^{d} u^{\{i\}} \bigoplus_{i=1}^{d} \bigoplus_{i<j} u^{\{i,j\}} \ldots$$

ANOVA and Factor Graph





Density Estimation using Sparse Grids

$$\tilde{p} = \arg\min_{f \in V} \int_{\Omega} (f(x) - p_{\epsilon}(x))^2 dx + \lambda \left| |\Lambda f| \right|_{L^2}^2$$

where f belongs to an exponential family.

$$\tilde{p}(x) = \frac{exp(u(x))}{\int exp\big(u(z)\big)dz} dx$$

$$u(x) = \sum_{j=1}^{N} \alpha_j \phi_j(x)$$

N: Number of Grid Points
α: Coefficients
$\phi(x)$: Basis functions

Density Estimation using Sparse Grids

$$A\alpha^{i+1} = E_{empir} - \Phi(\alpha^i)$$

$$E_{empir} = \frac{1}{M}\sum_{i=1}^{M} \phi_k(x_i)$$

$$(A)_{k,j} = a(\phi_k, \phi_j) = \lambda \int \nabla\, \phi_k(x) \nabla \phi_j(x) dx$$

$$\Phi(\alpha^i) = \int \phi_k(x)\, \widetilde{p}(x) dx$$

$\boldsymbol{while}\ ||A\alpha^{i+1} - E_{empir} + \Phi(\alpha^{i+1})|| > \epsilon\ and\ i < i_{max}\ \boldsymbol{do}$

$\quad Solve\ A\widetilde{\alpha} = E_{empir} - \Phi(\alpha^i)$

$\quad \alpha^{i+1} = \alpha^i + \omega\widetilde{\alpha}$

$\quad Compute\ \Phi(\alpha^{i+1})$

$\boldsymbol{end}$

Density Estimation using Sparse Grids

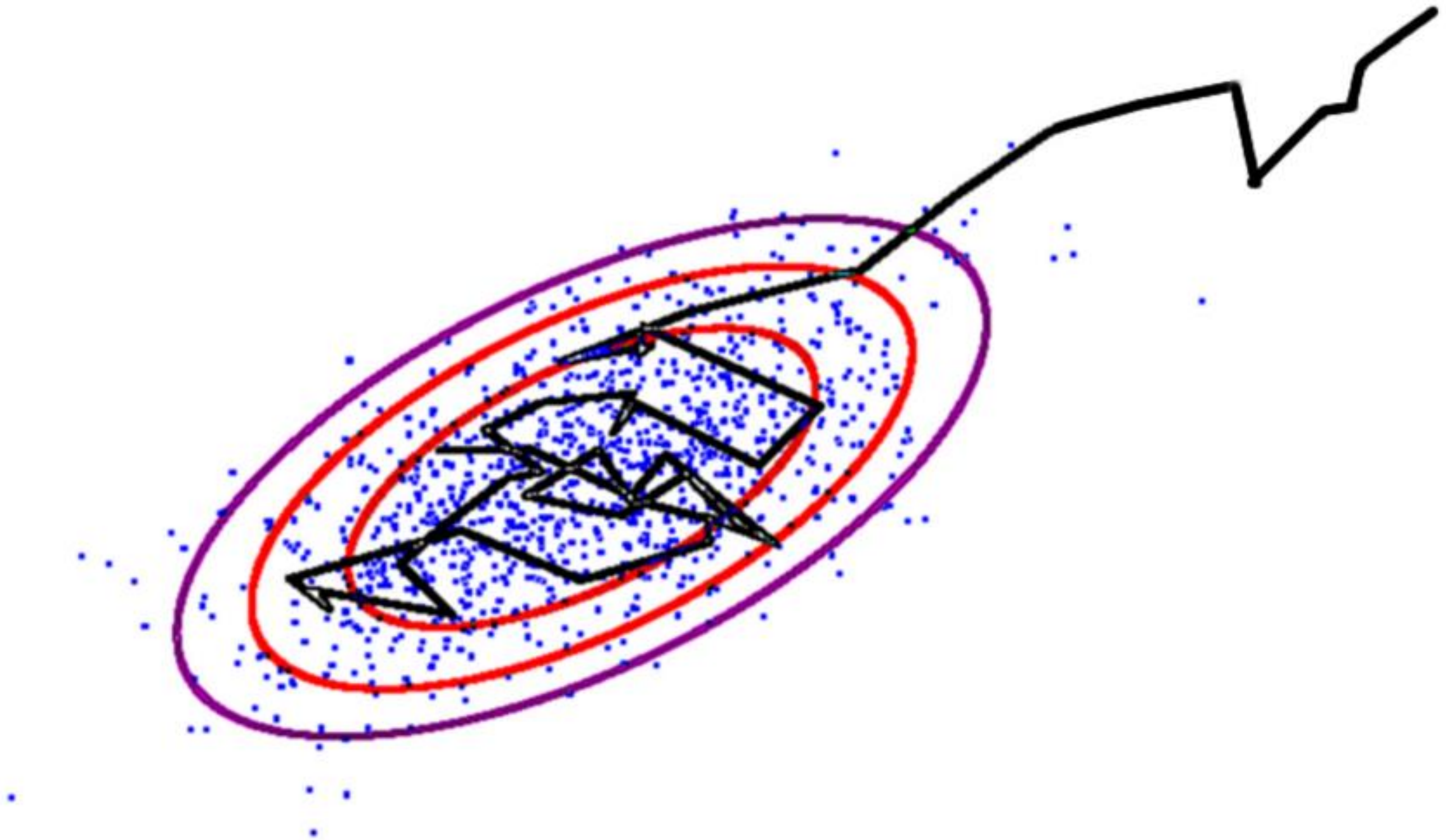
Computing the Expected Value

$$\Phi\left(\alpha^i\right) = \int \phi_k(x) \;\; \frac{exp(\sum_{j=1}^{N} \alpha_j^i \phi_j(x))}{\int exp(\sum_{j=1}^{N} \alpha_j^i \phi_j(z)) dz} \, dx$$

$$E\left[\phi_k(x)\right] = \int \phi_k(x) \;\; \tilde{p}(x) \;\; dx$$

- ✓ Monte Carlo Integration
- ✓ Markov Chain Monte Carlo

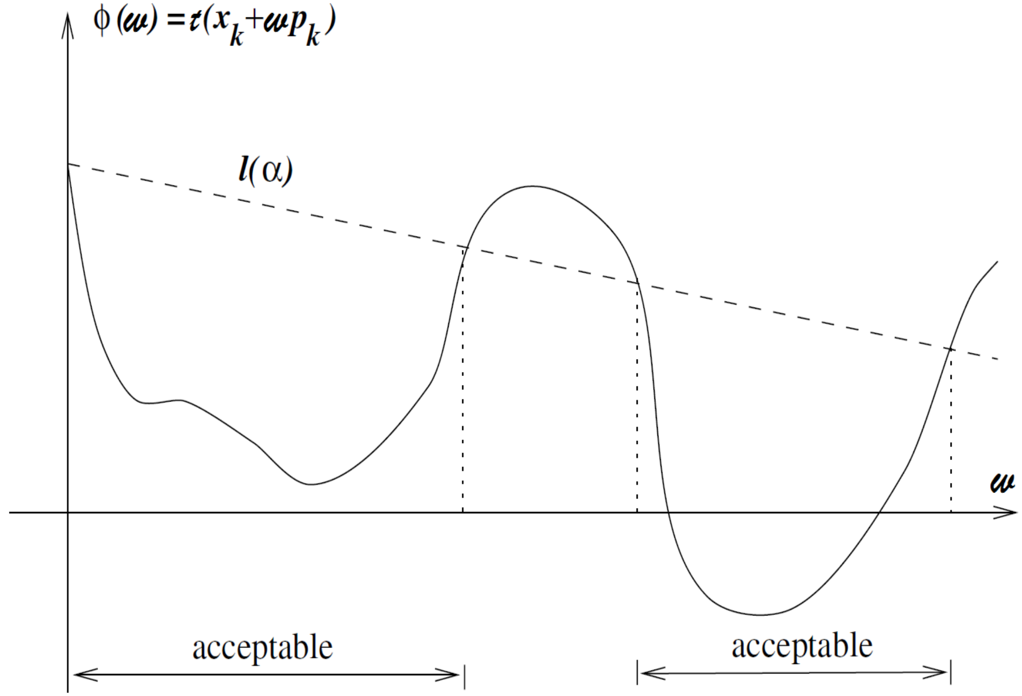
Markov Chain Monte Carlo



Source: http://www.juergenwiki.de/work/wiki/doku.php?id=public:mcmc

Performance Improvement Strategies

Dynamic Learning Rate using Armijo Line Search with pseudo likelihood

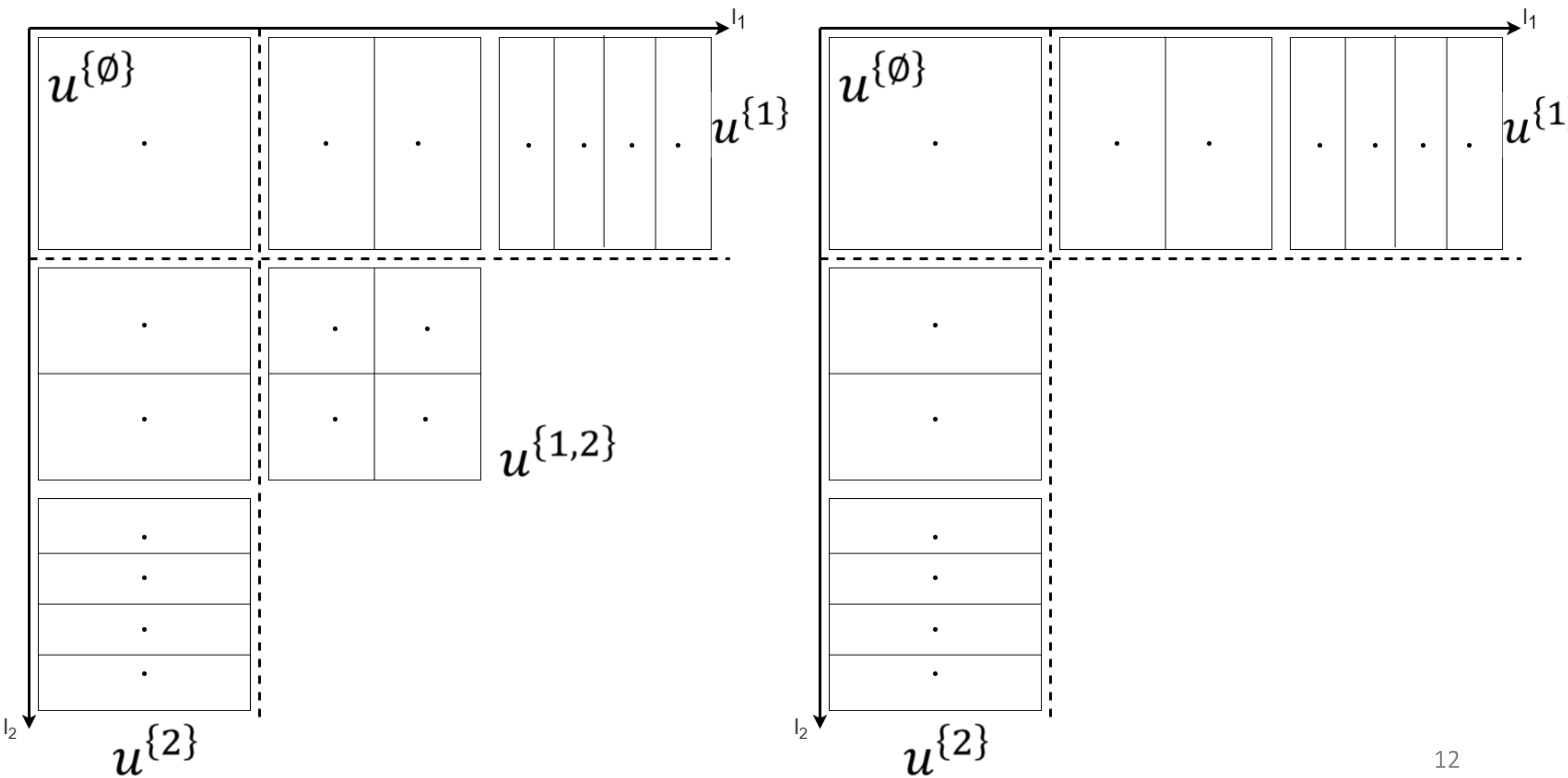


Source: J. Nocedal and S. Wright. Numerical Optimization. 2006.

Grid Coarsening

✓ Mean coefficient/volume thresholding

$$\boldsymbol{if}\ mean\ of\ |alphas\ of\ factor| < \alpha_{threshold}\ \boldsymbol{then}$$
$$add\ factor\ to\ delete\ list$$



Algorithm

$\boldsymbol{while}\ ||\nabla L|| > \ \epsilon\ and\ i < i_{max}\ \boldsymbol{do}$

$Solve\ A\tilde{\alpha}\ = E_{empir} - \Phi(\alpha^i) - A\alpha^i$

$Choose\ \omega\ using\ Armijo\ line\ search$

$\alpha^{i+1} = \alpha^i + \omega\tilde{\alpha}$

$Compute\ \Phi(\alpha^{i+1})$

$Compute\ \nabla L = E_{empir} - \Phi(\alpha^{i+1}) - A\alpha^{i+1}$

$grid, \alpha = grid\_coarsening(grid, \alpha, factor\ graph)$

$\boldsymbol{if}\ grid\ updated\ \boldsymbol{then}$

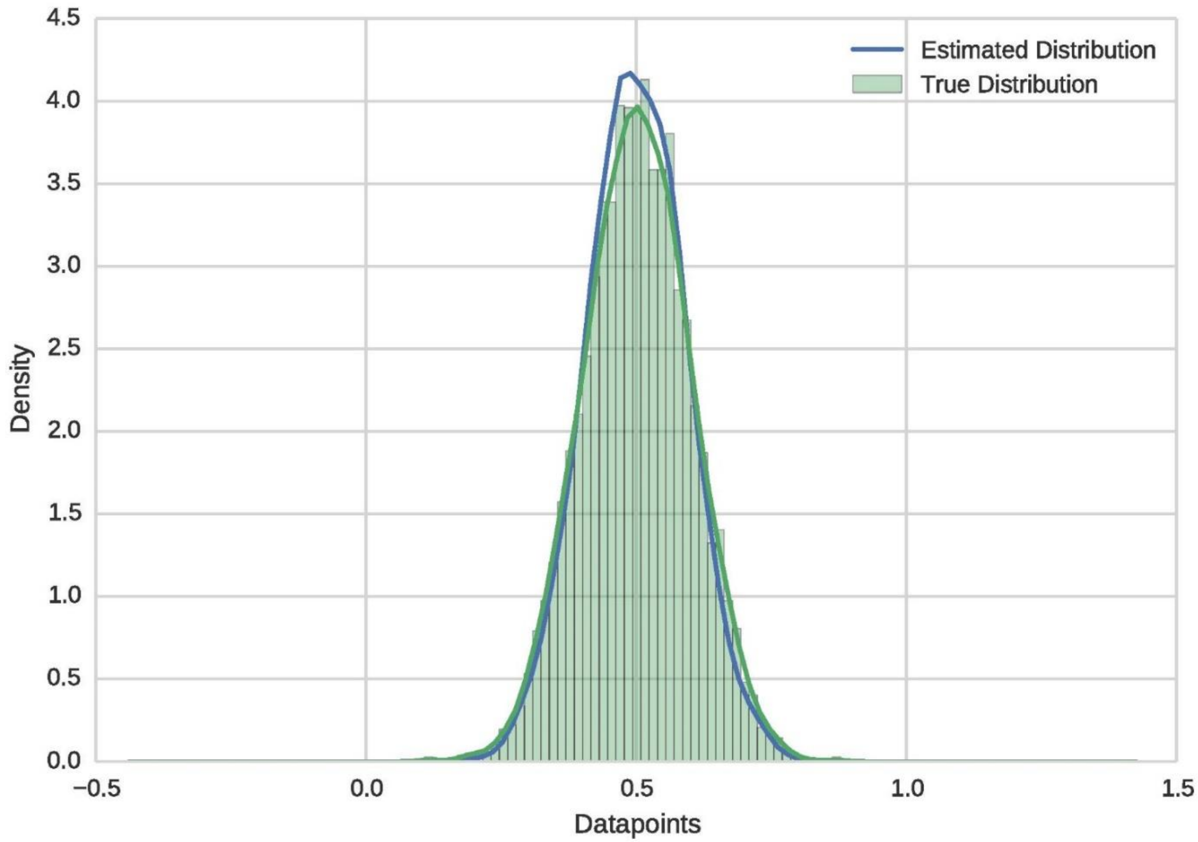
$Compute\ A, E_{empir}, \Phi(\alpha^i) and\ \nabla L$

$\boldsymbol{end}$

Experiment and Results

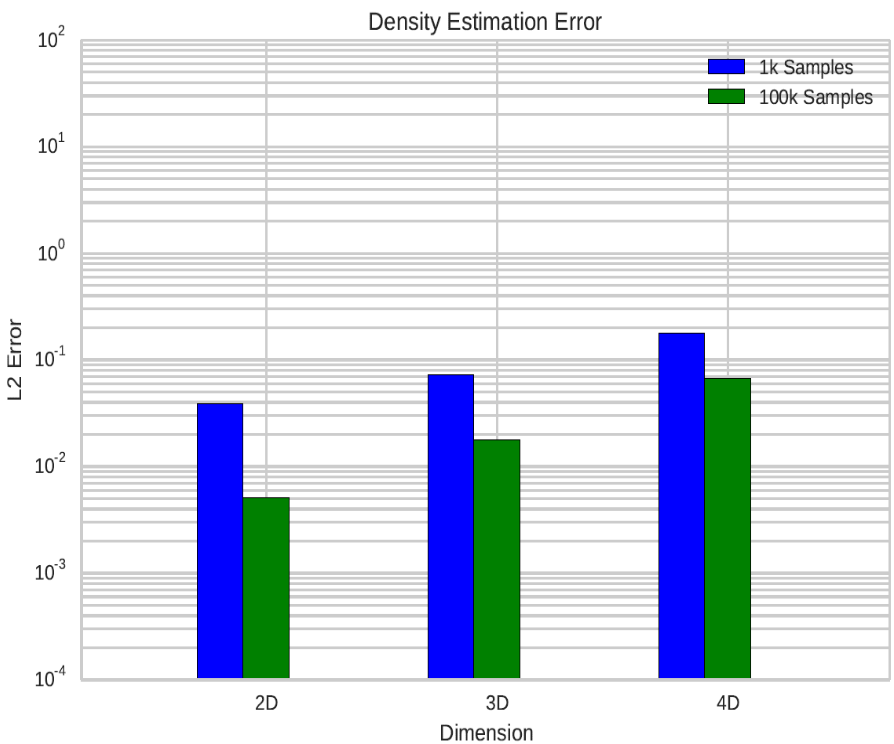
Datasets are sampled in each dimension independently from normal distribution with mean 0.5 and standard deviation 0.1.

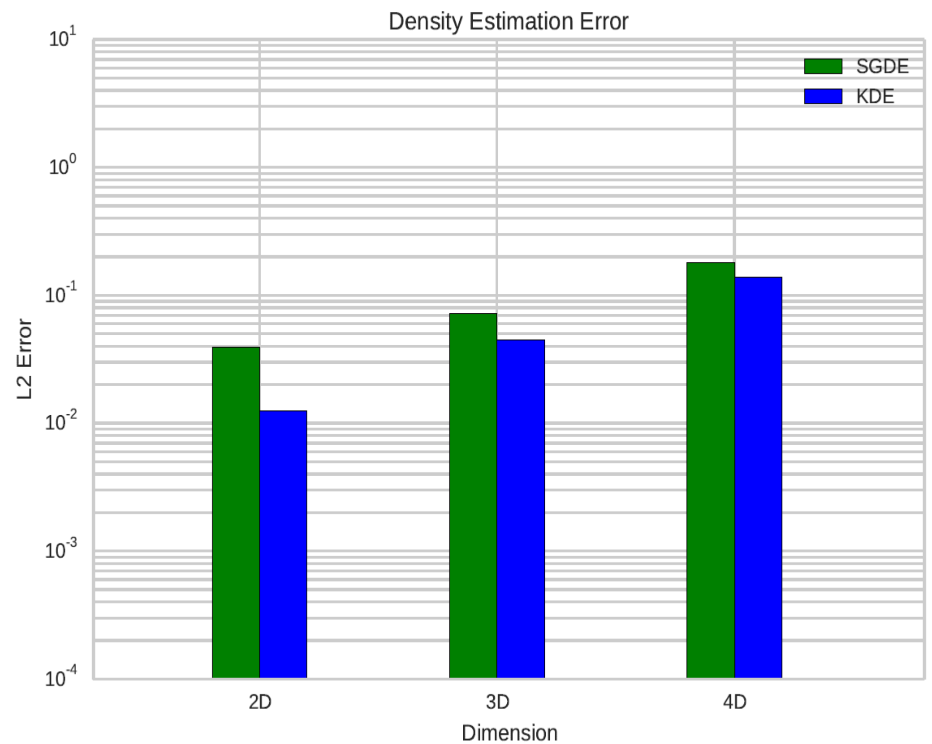
KS 2-sample test returned a p-value > 70%



Experiment and Results

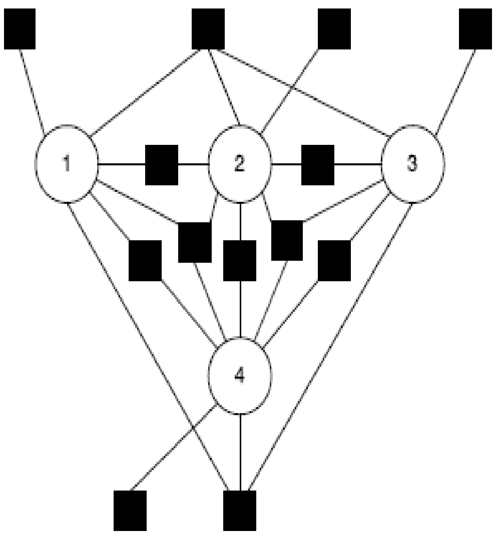
$$\frac{1}{M}\sqrt{\sum_{x \in D}\left(p(x)-\hat{p}(x)\right)^2}$$



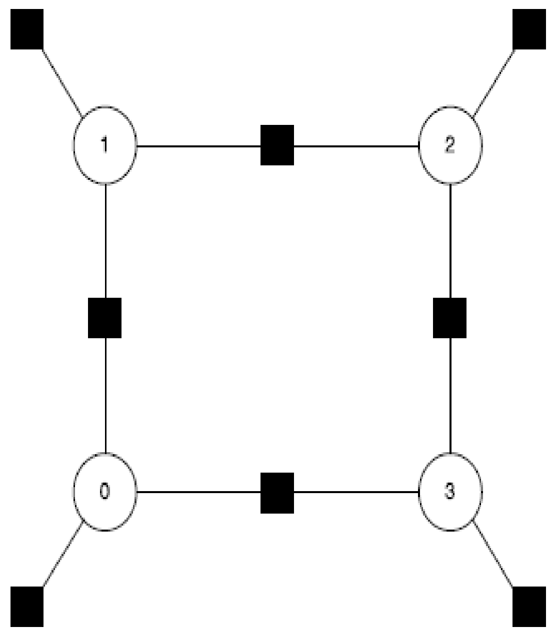


Experiment and Results

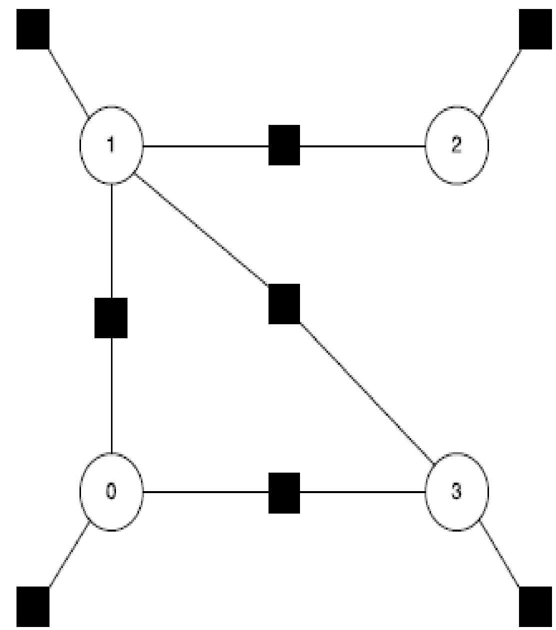
Datasets are sampled based on a particular structure, defined in terms of dependencies of the variables.



(a) Initial Structure

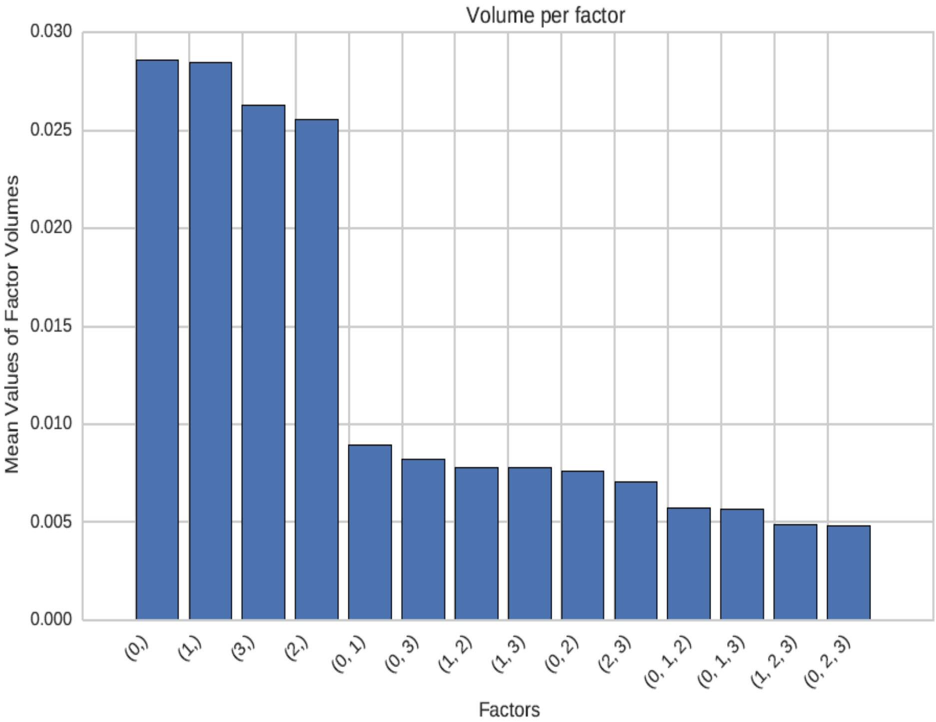


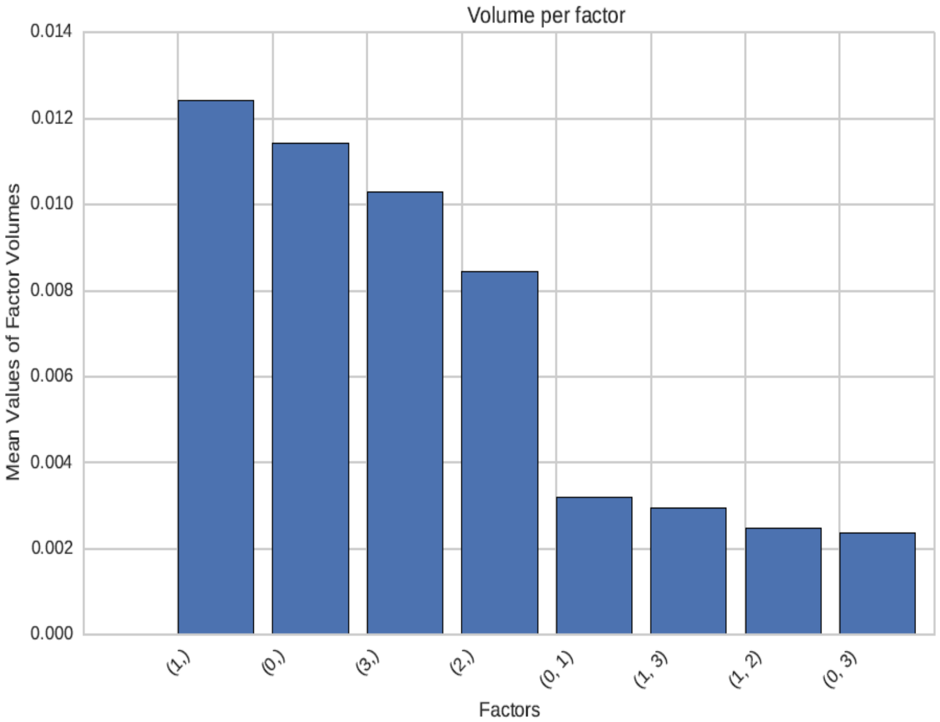
(b) Actual Structure



(c) Predicted Structure

Experiment and Results





References

[1] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1986.

[2] Hans-Joachim Bungartz and Michael Griebel. *Sparse grids*. Acta Numerica, 13:1123, 2004.

[3] P. Hahnen. *Nichtlineare numerische Verfahren zur multivariaten Dichtesch ätzung*. Diplomarbeit, Institut für Numerische Simulation, Universität Bonn, November 2006.

[4] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2006.

[5] Benjamin Peherstorfer. *Model Order Reduction of Parametrized Systems with Sparse Grid Learning Techniques*. Dissertation, Department of Informatics, Technische Universität München, October 2013.

[19] Sebastian Soyer. *Nonlinear density estimation with applications in astronomy*. 2014.