Constructing Shrinkage Priors in High Dimensions

Natesh S. Pillai
Department of Statistics,
Harvard University

July 25, 2013
SAHD
Duke University, NC

Collaborators

- Anirban Bhattacharya (Texas A&M)
- Debdeep Pati (FSU)
- David B. Dunson (Duke)

Bayes in High Dimensions

- Bayesian Estimation in "Large-p, small-n".
- Statistical Efficiency *vs.* Computational Efficiency, a key issue.
- In this talk: a concrete formulation of this problem.

Conversation with Peter E. Huybers

- Motivation: Time variability in covariance patterns for climate data: stationarity?
- Instrumental measurements, only for the past $n = 150$ years.
- Measurements on $p = 2000$ latitude-longitude points.
- Estimate $O(p^2)$ parameters.
- Need judicious modeling.

Motivation

- An important class of models: Latent factor methods (West, 2003; Lucas et al., 2006; Carvalho et al., 2008).
- Set $y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}, i = 1, \ldots, n$
- $y_i \sim \mathrm{N}_p(0, \Sigma)$
- Goal: Estimate Σ.
- Note $p \gg n$.

Gaussian factor models

- Unstructured Σ has $O(p^2)$ free elements
- Assume a factor model

$$\Sigma = \Lambda\Lambda' + \sigma^2 \mathrm{I}_p$$

  via parsimonious factorization
- $k = O(1)$, the number of factors.
- Λ is the factor loadings.
- Λ is $p \times k$ and thus model complexity $O(p)$ - huge dimensionality reduction, but still challenging.

Sparse factor modeling

- *Sparse factor modeling* (West, 2003); also (Lucas et al., 2006; Carvalho et al., 2008) and many others
- Allow zeros in loadings.
- Assume each column of Λ has only s non-zero elements.
- Here s denotes the sparsity.

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)
- Efficient Estimators based on Thresholding:

$$\hat{\Sigma}_{ij} = \Sigma_{ij}^{\text{sample}} 1_{|\Sigma_{ij}^{\text{sample}}| > t_n} \ .$$

High-dimensional covariance estimation

- 'Frequentist' solution – MLE doesn't work.
- Start with sample covariance matrix:

$$\Sigma^{\text{sample}} = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^T \ .$$

- Great interest in regularized estimation (Bickel & Levina, 2008a, b; Wu and Pourahmadi, 2010, Cai and Zhou, 2011 ...)
- Efficient Estimators based on Thresholding:

$$\hat{\Sigma}_{ij} = \Sigma_{ij}^{\text{sample}} 1_{|\Sigma_{ij}^{\text{sample}}| > t_n} \ .$$

- Unstable; Confidence intervals..?

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?
- Successfully exploited in "classical statistics": *i.e.*, *fixed-p*, large-n situation.

Bayes-Frequentist Synergy

- Question: Given most regularization estimators are posterior modes of a Bayesian model, can one run a Markov Chain Monte Carlo algorithm to sample from the posterior distribution, and compute the uncertainty intervals?
- Successfully exploited in "classical statistics": *i.e.*, *fixed-p*, large-n situation.
- Here we assume, $p_n = O(e^{n^\alpha})$ with $\alpha < 1/3$ (ultra high-dimensions).

Bayes-Frequentist Synergy

- Set $k = 1$, thus

$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- Questions:
  1. What is the minimax rate for estimating Σ?
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Bayes-Frequentist Synergy

- Set $k = 1$, thus
$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$
- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- Questions:
  1. What is the minimax rate for estimating Σ?
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Answer to the above two questions: a first step towards Bayes-Frequentist agreement in this "large-p, small-n" problem.

Assumptions:

- Recall, $k = 1$, thus

$$\Sigma = \sigma^2 \mathrm{I}_p + \Lambda\Lambda'$$

- Λ is a $p \times 1$ vector, with only s many non-zeroes.
- $p_n = O(e^{n^\alpha})$ with $\alpha < 1/3$
- Key facet:

$$\sigma^2 < \|\Lambda\Lambda'\|_2 = \|\Lambda\|^2 = O(\log p_n)$$

- Thus Σ is not a "small" perturbation of identity (different from other common assumptions...)

Minimax rate:

- **Theorem (Minimax Lower Bound)**

  (Pati, Bhattacharya, P., Dunson, 2013)

$$\inf_{\hat{\Sigma}} \sup_{\Sigma} \|\hat{\Sigma} - \Sigma\|_2 \geq \sqrt{\frac{(\log p_n)^3 s}{n}}$$

  - Proof uses a variant of Le Cam's method/ Fano's Lemma.
  - Questions:
    1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

- Where to look for possible priors?

Posterior rate

- What prior on the vector Λ leads to a posterior which concentrates at the minimax rate?
- Let Σ_0 be the true data generating parameter.
- We seek ϵ_n such that

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

- Where to look for possible priors?
- First choice: point mass priors. These can be thought of as the Bayesian analogue of thresholding estimates.

Posterior rate: Point Mass Priors

- Set

$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$

where $g(\cdot)$ has exponential or heavier tails.

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.
- If s is unknown, set a hyper-prior (Scott & Berger 2010, Castillo & van der Vaart, 2012)
$$\pi \sim \text{Beta}(1, p+1).$$

Posterior rate: Point Mass Priors

- Set
$$\Lambda_j \sim (1-\pi)\delta_0 + \pi g(\cdot)$$
where $g(\cdot)$ has exponential or heavier tails.
- If s is known, then $\pi = s/p$ is a natural choice.
- If s is unknown, set a hyper-prior (Scott & Berger 2010, Castillo & van der Vaart, 2012)
$$\pi \sim \text{Beta}(1, p+1).$$
- Has connections to automatic multiplicity adjustments, and also optimal in other contexts.

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- **Theorem (Posterior Convergence Rate)**

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \text{Data}) = 0.$$

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).
  - The sample covariance will not be efficient in detecting the points.

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \text{Data}) = 0.$$

  - Proof involves novel ideas, and uses results from non-asymptotic random matrix theory (Vershynin 2010, Tropp 2012).
  - The sample covariance will not be efficient in detecting the points.
  - Take the low dimensional projections of the data, and then compute the sample covariance matrix.

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Theorem (Posterior Convergence Rate)

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \mathrm{Data}) = 0.$$

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- **Theorem (Posterior Convergence Rate)**

  For point mass priors, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \text{Data}) = 0.$$

  - Questions:
    1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? $=$ Point mass priors achieve this!

Have we solved the problem?

- ...Not yet!

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.
- Effective sample size is small; Point mass priors are statistically efficient, but computationally NOT efficient!

Have we solved the problem?

- ...Not yet!
- The MCMC algorithm for sampling the posterior has to explore a model space of dimension $O(2^p)$.
- Curse of dimensionality catches up fast; not even feasible for moderate p.
- Effective sample size is small; Point mass priors are statistically efficient, but computationally NOT efficient!
- OK, what now?

Funny you should ask...

- Continuous Shrinkage Priors!

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)
- Polson & Scott (2010) unifies them as

$$\Lambda_j \sim N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Funny you should ask...

- Continuous Shrinkage Priors!
- Appealing computationally & philosophically to relax assumption of exact zeros.
- Zillions of them (Park and Casella, 2008; Carvalho, Polson and Scott, 2010; Armagan, Dunson and Lee, 2011; Hans, 2011,..)
- Polson & Scott (2010) unifies them as

$$\Lambda_j \sim N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- Many penalized least squares estimators correspond to mode of a Bayesian posterior (e.g., $L_1 \equiv$ Laplace prior)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)
- g inverse-gamma = (RVM, Tipping, 2001)

Global-local scale mixtures of Gaussians (Polson & Scott, 2010)

- Essentially all shrinkage priors can be represented as

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- τ - global shrinkage toward zero, ψ_j's - avoid over-shrinking signals locally
- g exponential = (Bayesian Lasso, Park & Casella, 2008; Hans, 2009)
- g inverse-gamma = (RVM, Tipping, 2001)
- g half-Cauchy = (Carvalho et al., 2009)

Posterior rate: Pati, Bhattacharya, P., Dunson, 2013

- **Theorem (Posterior Rate)**

  For most global-local shrinkage priors defined as above, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n \to \infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \text{Data}) \neq 0.$$

Posterior rate: Pati, Bhattacharya, P., Dunson, 2013

- **Theorem (Posterior Rate)**

  For most global-local shrinkage priors defined as above, with $\epsilon_n = \sqrt{\frac{(\log p_n)^3 s}{n}}$, *we have*

$$\lim_{n\to\infty} \mathbb{P}_{\Sigma_0}(\|\Sigma - \Sigma_0\|_2 \geq \epsilon_n | \text{Data}) \neq 0.$$

  - Questions:
    1. What is the minimax rate for estimating Σ?$= \sqrt{\frac{(\log p_n)^3 s}{n}}$
    2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? $=$ Point mass priors achieve this! Most global-local priors do NOT!

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- As before:

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. *A priori* independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- As before:

$$\Lambda_j \overset{ind}{\sim} N(0, \psi_j \tau), \quad \psi_j \overset{i.i.d}{\sim} g, \quad \tau \sim f$$

- Local scales ψ are apriori independent; thus no *a priori* borrowing of information across coordinates, needed for efficient shrinkage estimators!

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.

Statistical Inefficiency of Global-local priors

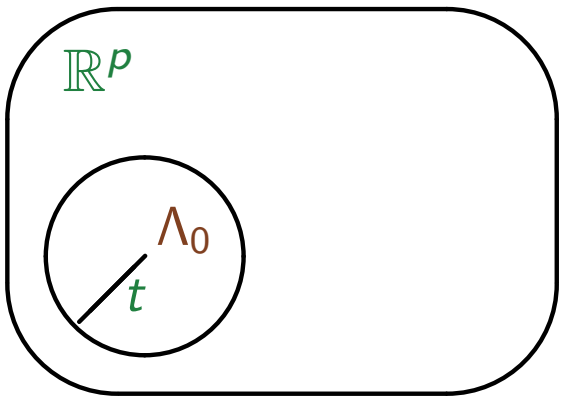
- What goes wrong? Two things:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- In constructing prior distributions, we have to make sure that the prior gives sufficient mass around the "true parameter".

Statistical Inefficiency of Global-local priors

- What goes wrong? Two things:
  1. A priori independence of coordinates: inefficient shrinkage, a la Stein.
  2. Concentration of Measure.
- In constructing prior distributions, we have to make sure that the prior gives sufficient mass around the "true parameter".
- Joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse Λ_0

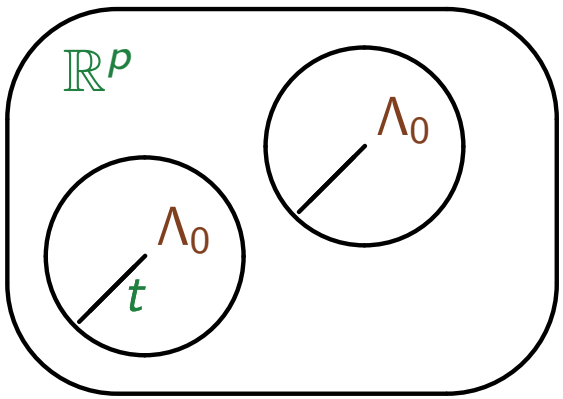
Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



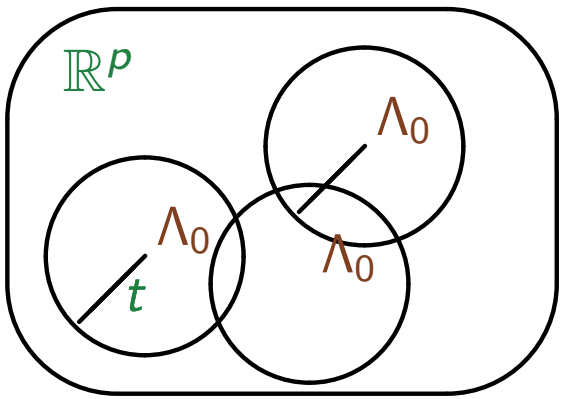
Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



Statistical Inefficiency of Global-local priors

- Need joint concentration $\mathbb{P}(\|\Lambda - \Lambda_0\|_2 \leq t)$ crucial for sparse θ_0



Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need non-asymptotic concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need non-asymptotic concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

- If Λ_j's are i.i.d. $N(0, 1)$, then

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-Cp}$$

Prior concentration - some initial examples

- Recall The truth $\Lambda_0 \in \mathbb{R}^p$: with at most s non-zero elements.
- Focus: Need non-asymptotic concentration bounds for

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p})$$

- If Λ_j's are i.i.d. $N(0,1)$, then

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-Cp}$$

- On the other hand, for suitable point mass priors (g Laplace)

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \geq e^{-Cs \log p}$$

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration
- Bayesian LASSO:

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-C\sqrt{p}}$$

Prior concentration

- KEY RESULT: Most continuous shrinkage priors give poor concentration
- Bayesian LASSO:

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \leq e^{-C\sqrt{p}}$$

- Thus the concentration improves only a little.

Dirichlet Laplace prior & properties

- Next key idea: Can we induce dependence across local scales?
- We propose a simple dependent modification leading to optimal concentration & efficient computation

$$\Lambda_j \sim \text{Double Exp}(\phi_j \tau)$$

- IDEA: Constrain ϕ to the simplex - this allows for dependence
- We let $\phi \sim \text{Diri}(\alpha, \ldots, \alpha)$ - $\alpha < 1$ favors small # dominant values with remaining ≈ 0
- For this prior distribution,

$$\mathbb{P}(\|\Lambda - \Lambda_0\|_2 < \sqrt{p}) \geq e^{-Cs \log p}$$

- Thus matches the point mass prior distributions!
- We also have an efficient Gibbs sampling algorithm for sampling!

Minimax rate: Pati, Bhattacharya, P., Dunson, 2013

- Questions:
  1. What is the minimax rate for estimating Σ? $= \sqrt{\frac{(\log p_n)^3 s}{n}}$
  2. What prior on the vector Λ leads to a posterior which concentrates at the minimax rate? = Point mass priors achieve this! The Dirichlet-Laplace Prior above also achieves this!

Now, have we solved the problem?

- Not completely!
- There are different MCMC algorithms for posterior sampling.
- The only commonly used measure so far is the "effective sample" size.
- Hard to get exact bounds theoretically for most examples!
- For example, here is a general purpose algorithm.

Hybrid Monte Carlo

- Algorithm from Physics, (Duane et. al. (1987))
- Based on Hamiltonian Dynamics, conservation of energy.
- Imagine sampling from a probability distribution proportional to $e^{-g(x)}$, with $x \in \mathbb{R}^d$, $d \gg 1$.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = -\nabla g(x)$$

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

that preserves total energy.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \ \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

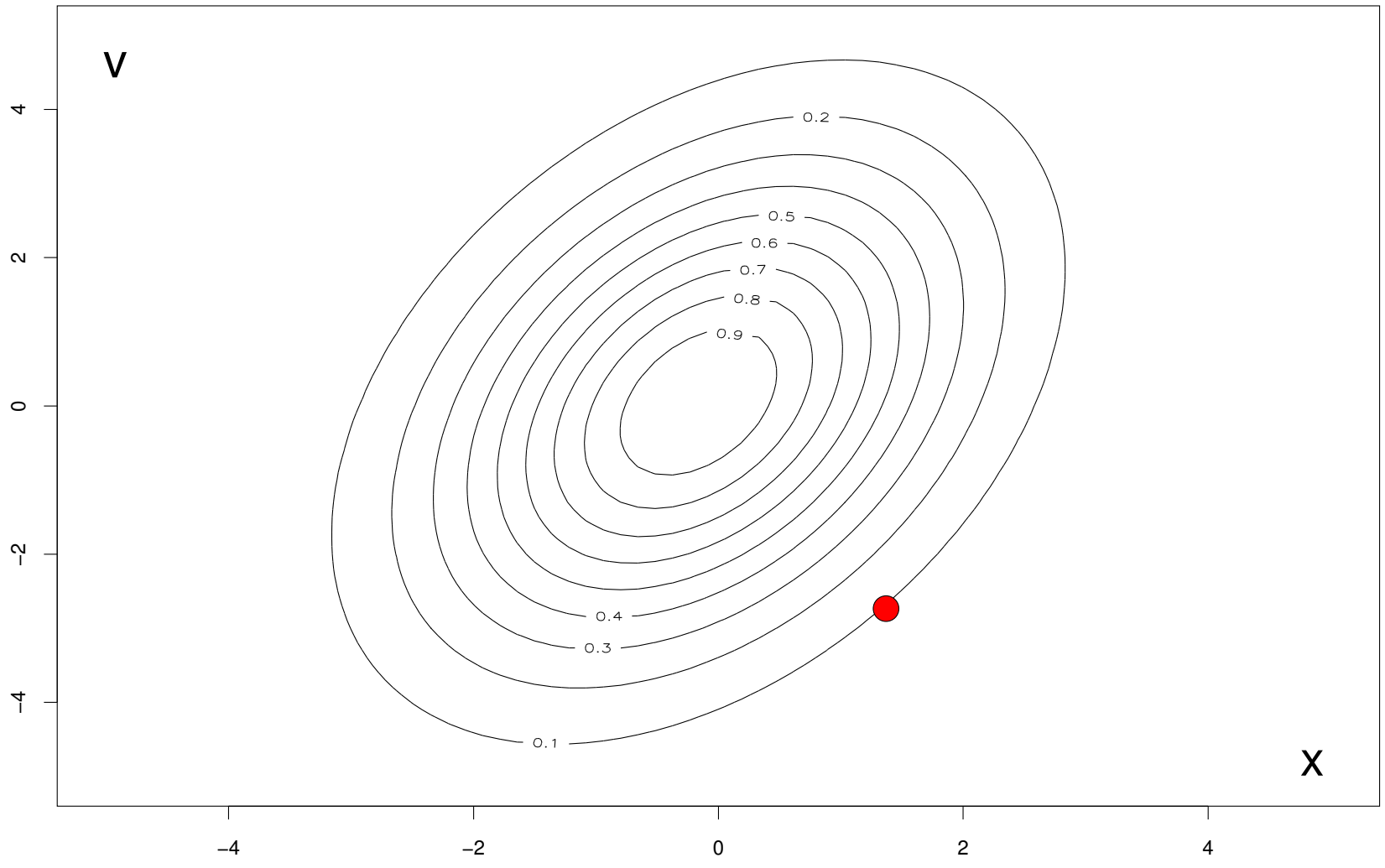
that preserves total energy.

- Equivalently the joint density

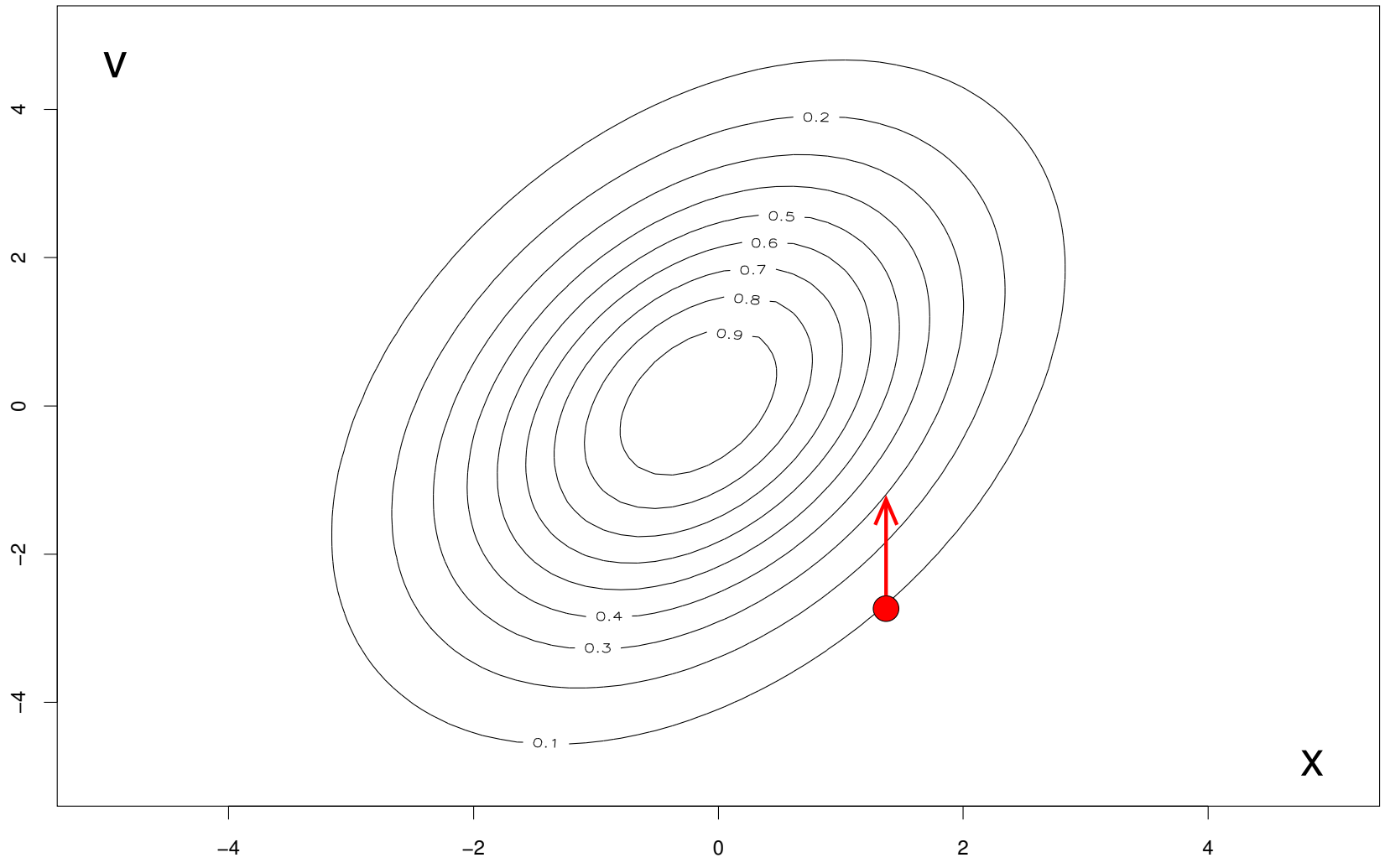
$$\exp\{-H(x, v)\} = \exp\{-g(x) - \frac{1}{2} v^2\}$$

is preserved.

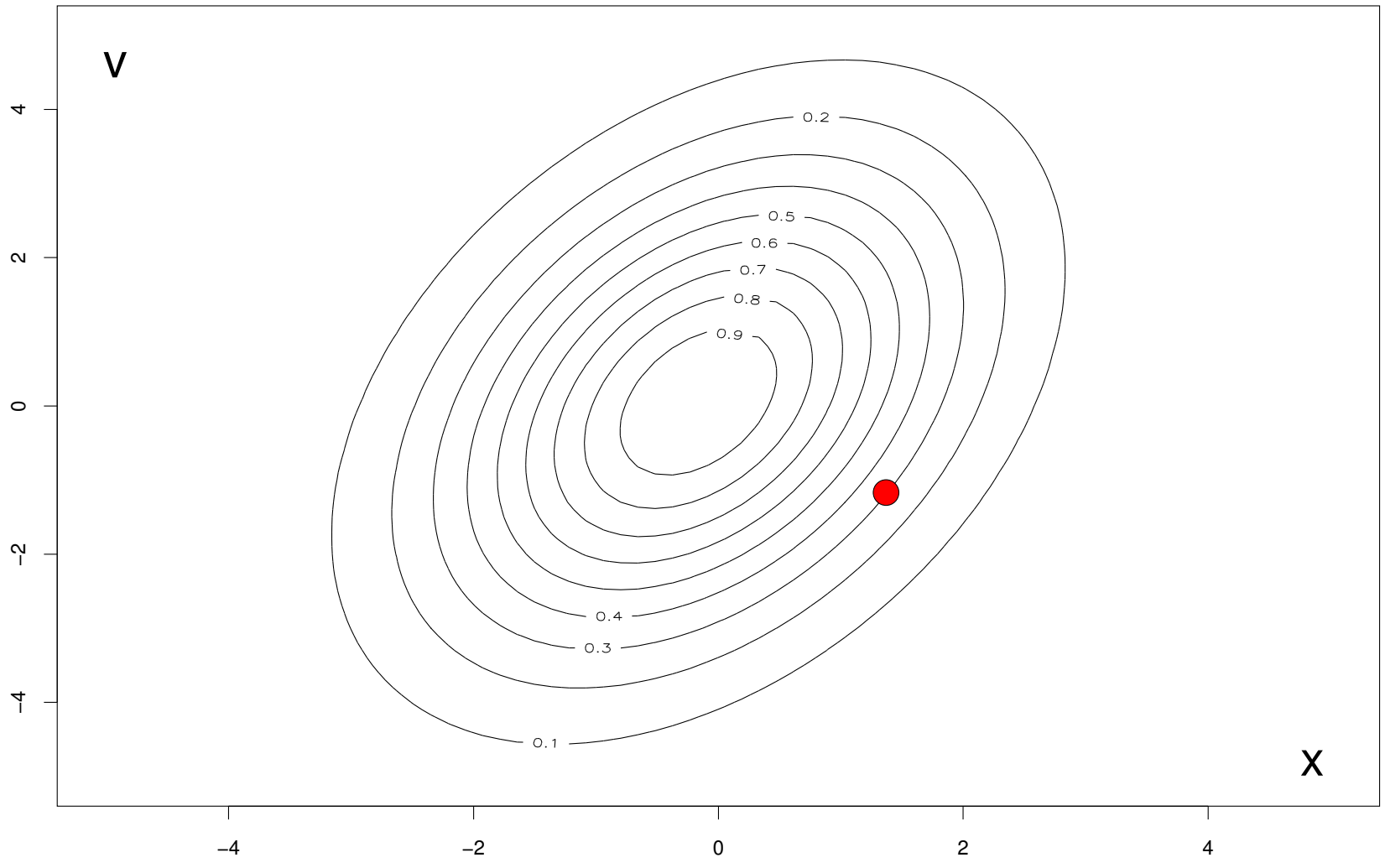
‘Exact’ Hamiltonian Dynamics



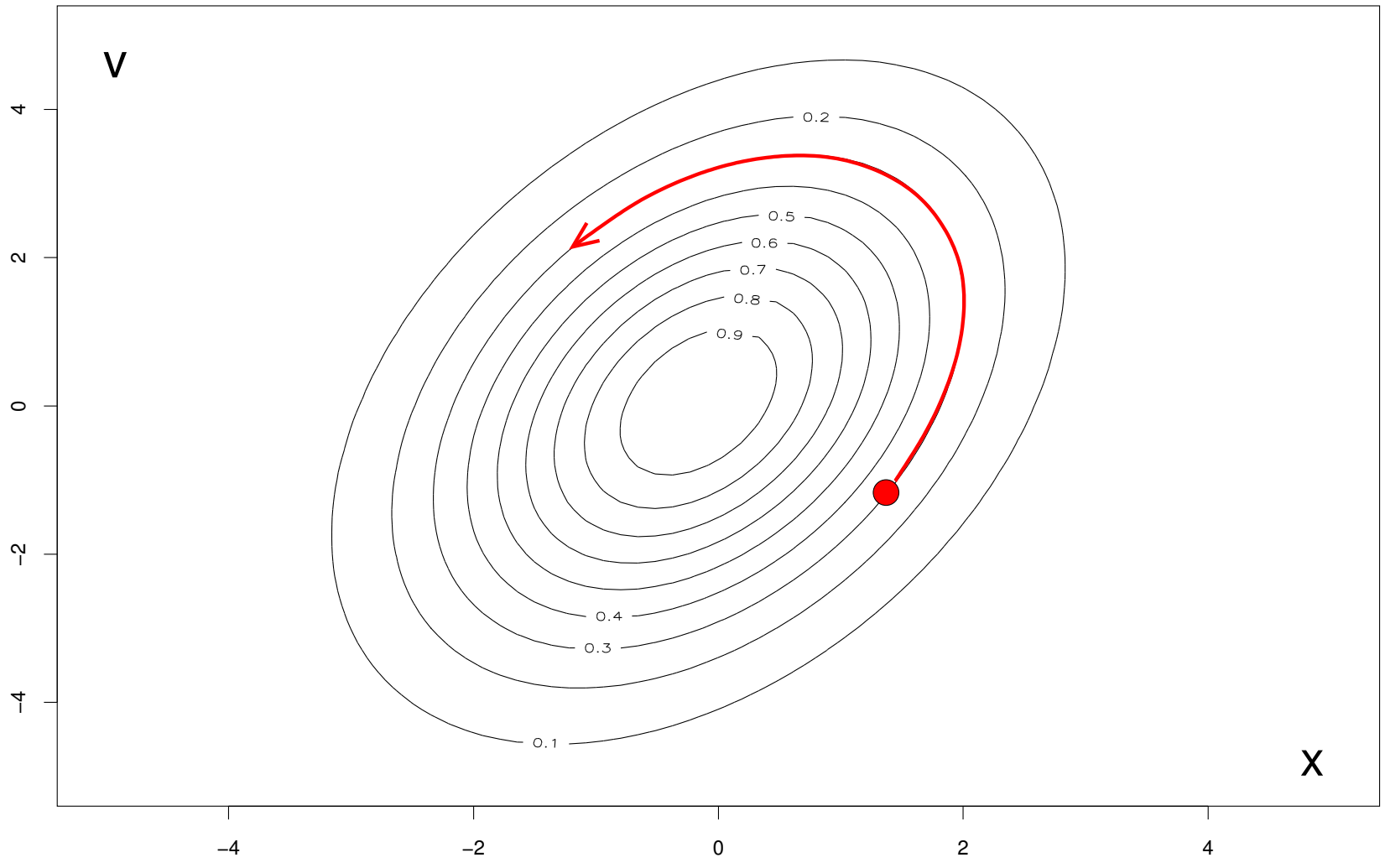
'Exact' Hamiltonian Dynamics



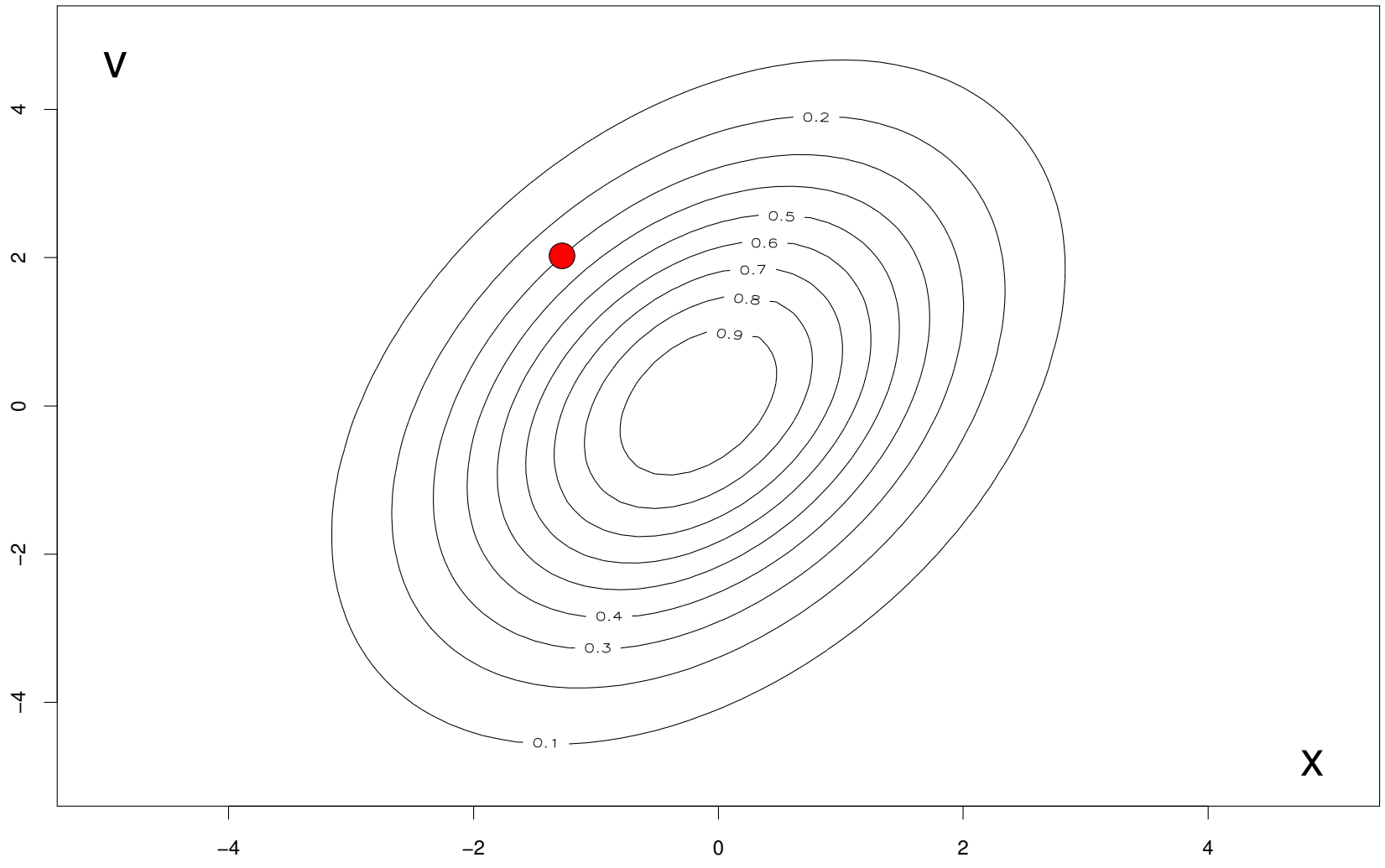
'Exact' Hamiltonian Dynamics



‘Exact’ Hamiltonian Dynamics



‘Exact’ Hamiltonian Dynamics



Paradigm Shift for Large Data Sets

- The computational complexity of the some of the commonly used MCMC algorithms are exponential in the data set.
- Need new theoretical framework for evaluating the efficiency MCMC algorithms with fixed computational complexity.
- Evaluate the efficiency of MCMC algorithms keeping the CPU time fixed- Widely open area!

Summary

- Concrete formulation of the statistical efficiency vs. computational efficiency.
- Under mild conditions, efficient posterior convergence is possible even if $p \gg n$.
- Prior concentration very important - should give enough probability near sparse subspaces.
- Appropriate point mass mixture priors can achieve this - prior probability of subset size important
- Most continuous shrinkage priors do not achieve this.
- Also developed a continuous shrinkage prior which does indeed meets both the theoretical and computational efficiency criteria.
- Other models, algorithms, ...

References

- Bayesian Shrinkage
- Posterior Contraction Rates in Sparse Bayesian Models for Massive Covariance Matrices
- Anirban Bhattacharya (Texas A&M), Debdeep Pati (FSU), David Dunson (Duke)

Acknowledgements

Thank you!