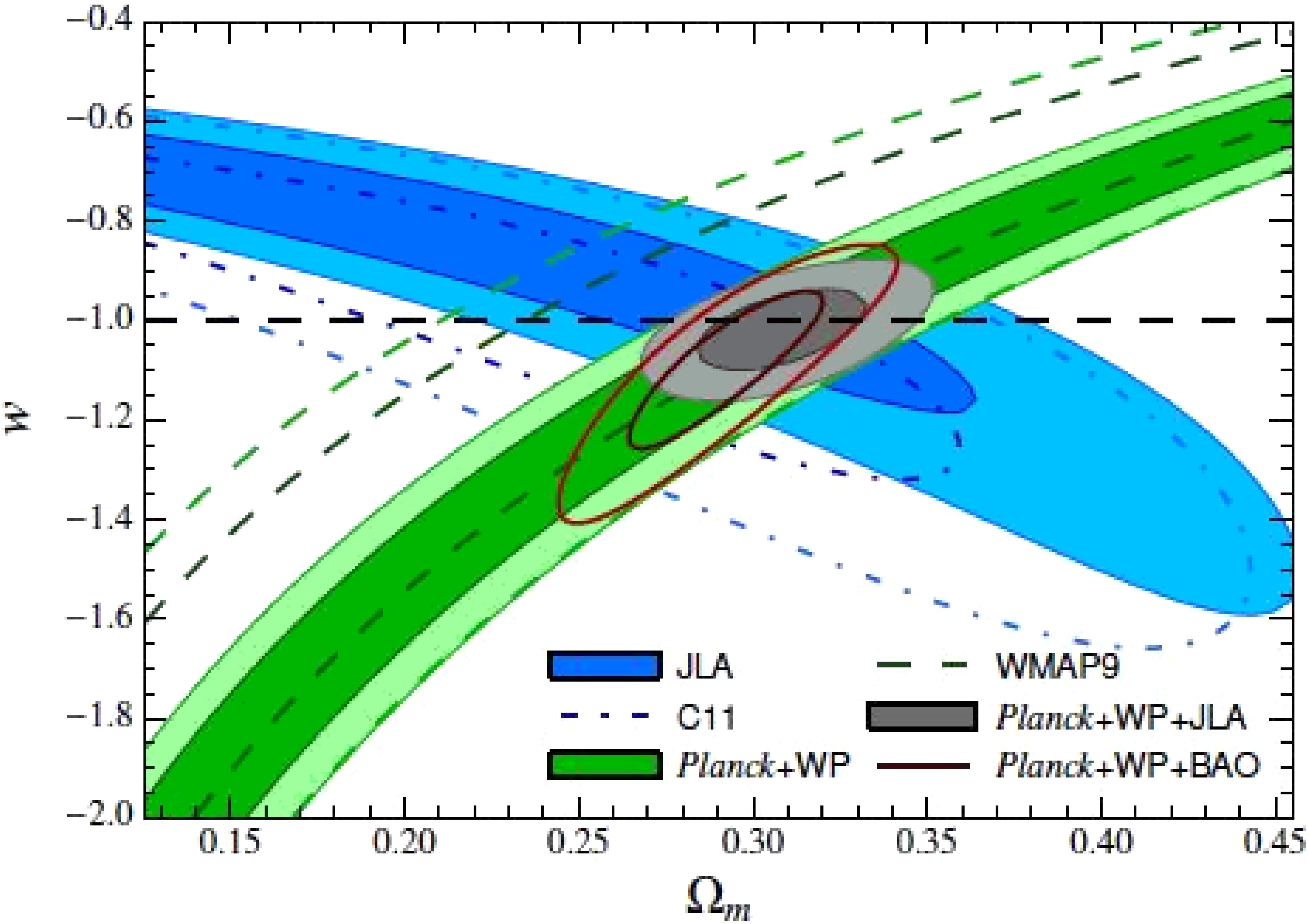
Learning priors, likelihoods, or posteriors

Iain Murray
School of Informatics, University of Edinburgh

Posteriors in Cosmology



https://bitbucket.org/joezuntz/cosmosis/

"Within the field of approximate Bayesian inference, **variational** *and* **Monte Carlo** *methods are currently the mainstay techniques."*

— `http://approximateinference.org/`

The Statistician (1987) 36, pp. 247–249

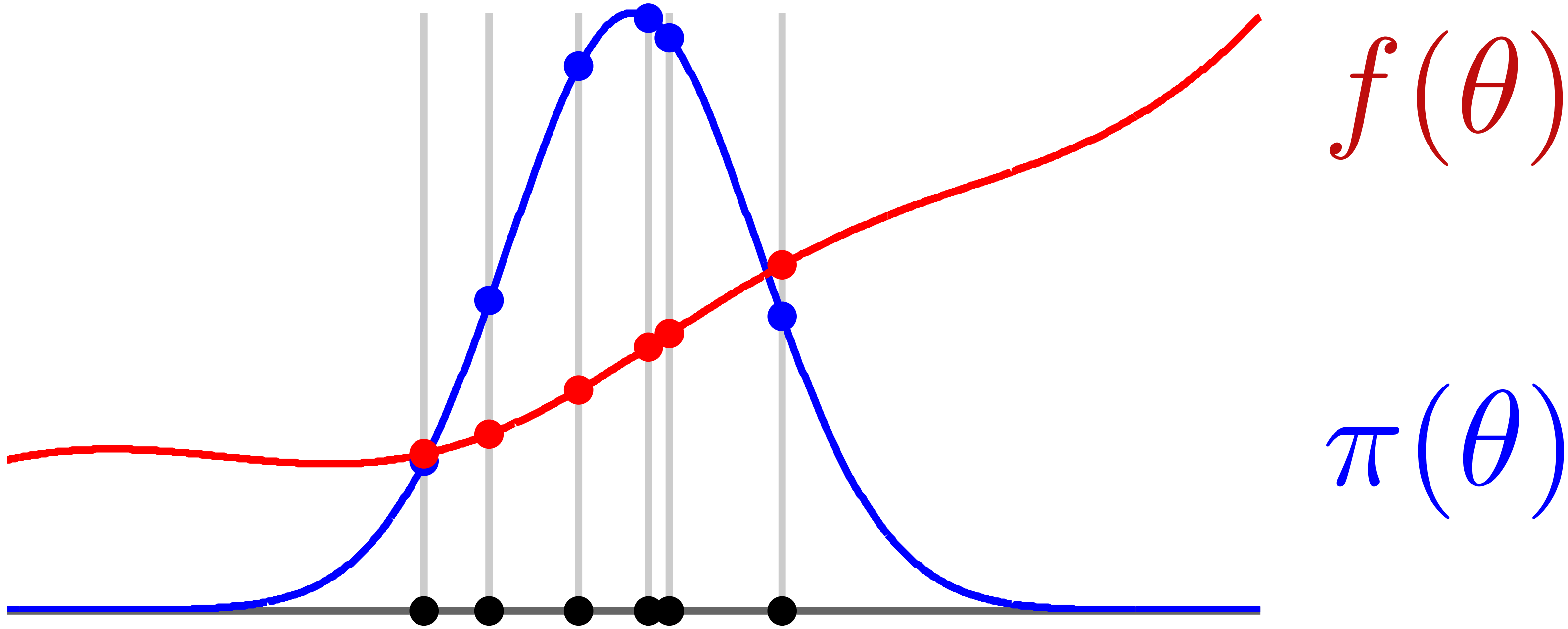
Monte Carlo is fundamentally unsound

A. O'HAGAN

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.

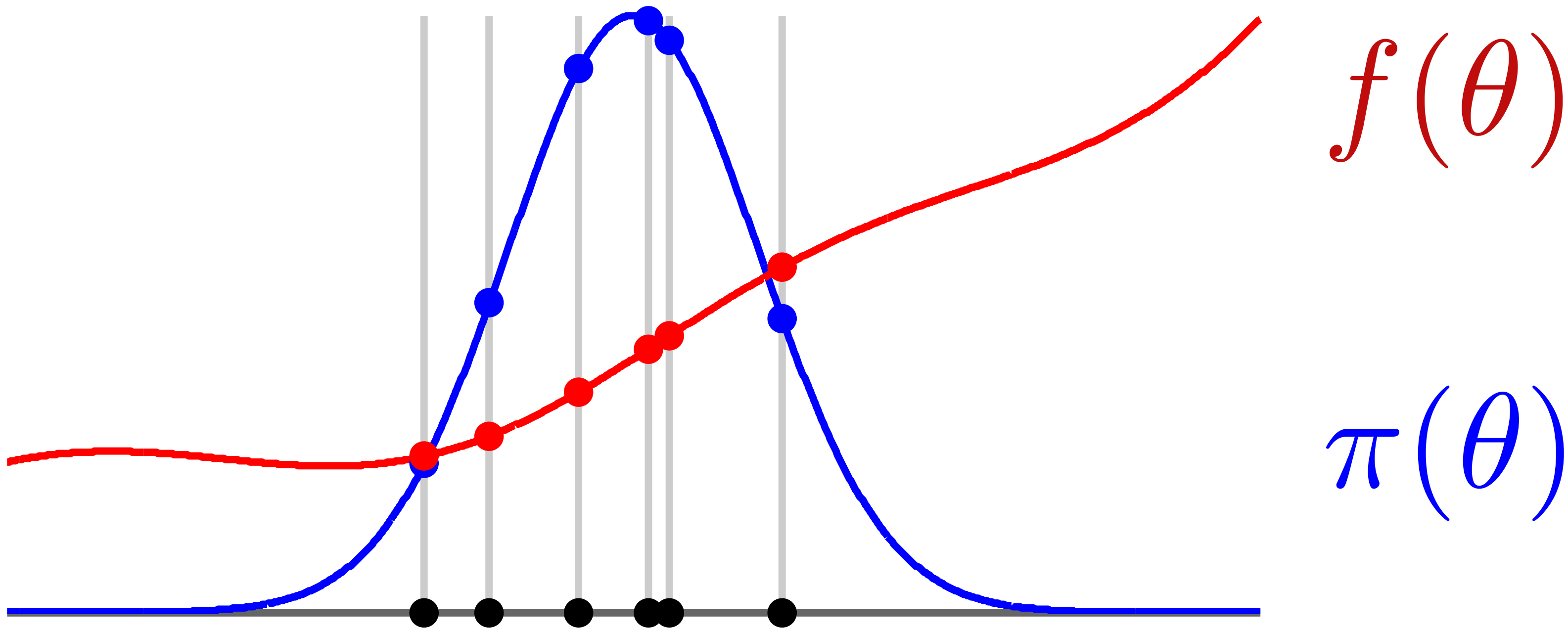
Abstract. We present some fundamental objections to the Monte Carlo method of numerical integration.

Why is Monte Carlo 'unsound'?



$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;\approx\; \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)}), \qquad \theta^{(s)} \sim \pi$$

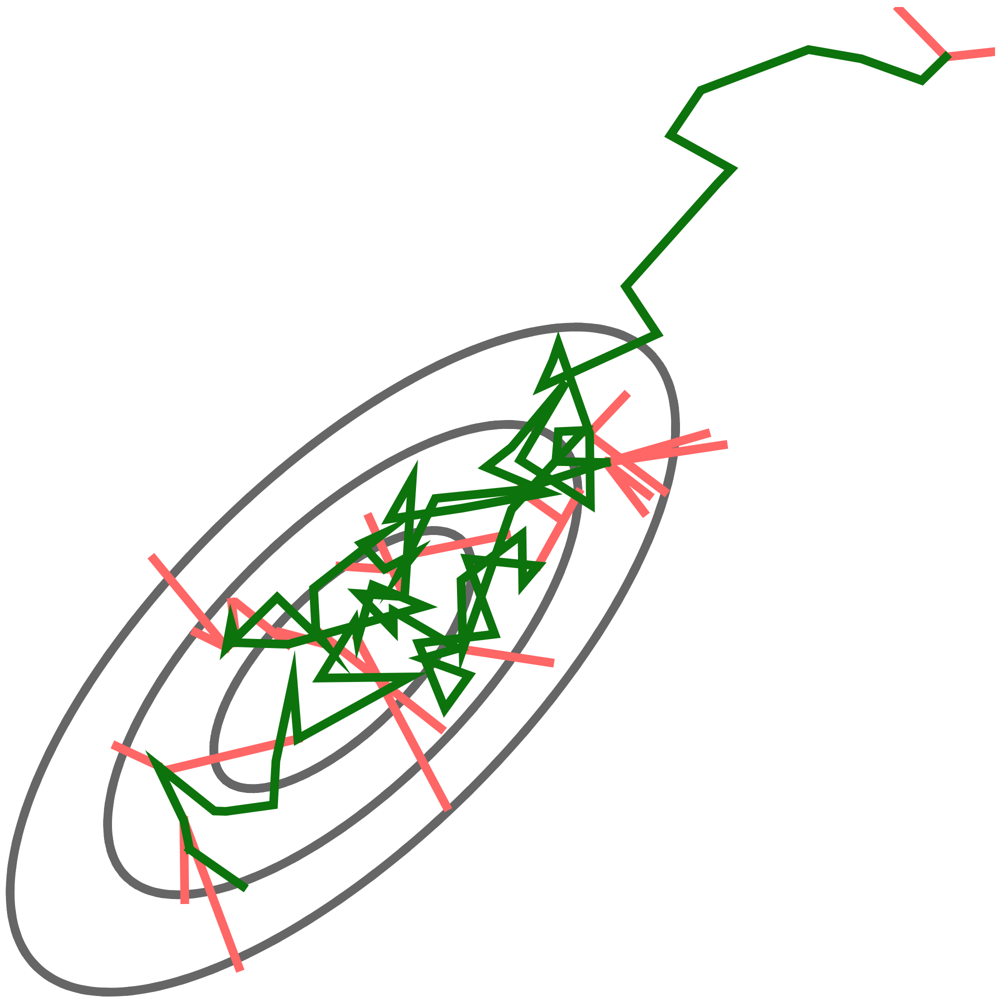
Why is Monte Carlo 'unsound'?



$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)}), \quad \theta^{(s)} \sim \pi$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})\,\frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}, \quad \theta^{(s)} \sim q$$

Metropolis–Hastings



$$\theta' \sim q(\theta'; \theta^{(s)})$$

```
if accept:
```

$$\theta \leftarrow \theta'$$

```
else:
```

$$\theta \leftarrow \theta^{(s)}$$

$$P(\text{accept}) = \min\left(1,\ \frac{p(\theta' \mid \mathcal{D})}{p(\theta^{(s)} \mid \mathcal{D})}\, \frac{q(\theta^{(s)}; \theta')}{q(\theta'; \theta^{(s)})}\right)$$

Recognition networks

$$\theta^{(s)} \sim p(\theta)$$

$$\mathbf{x}^{(s)} \sim p(\mathbf{x} \,|\, \theta^{(s)})$$

Training set: $\left\{\theta^{(s)}, \mathbf{x}^{(s)}\right\}_{s=1}^{S}$

Some of the relevant work

Hinton et al. (1995, Science) — Wake Sleep, Helmholtz machine

Morris (2001, UAI) — Recognition Networks

Blum & Francois (2010, S&C) — Conditional Gaussian, neural nets

Fan, Nott, Sisson (2012, Stat) — Mixture of experts

Mitrović, Dino Sejdinović, Teh (2016, ICML) — Kernel regression

...

Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation

Papamakarios and Murray (NIPS, 2016)
Lueckmann et al. (NIPS, 2017)

— Fit $\hat{p}(\theta \,|\, \mathbf{x})$ maximize $\sum_s \log \hat{p}(\theta^{(s)} \,|\, \mathbf{x}^{(s)})$

Mixture Density Networks (Bishop, 1994)

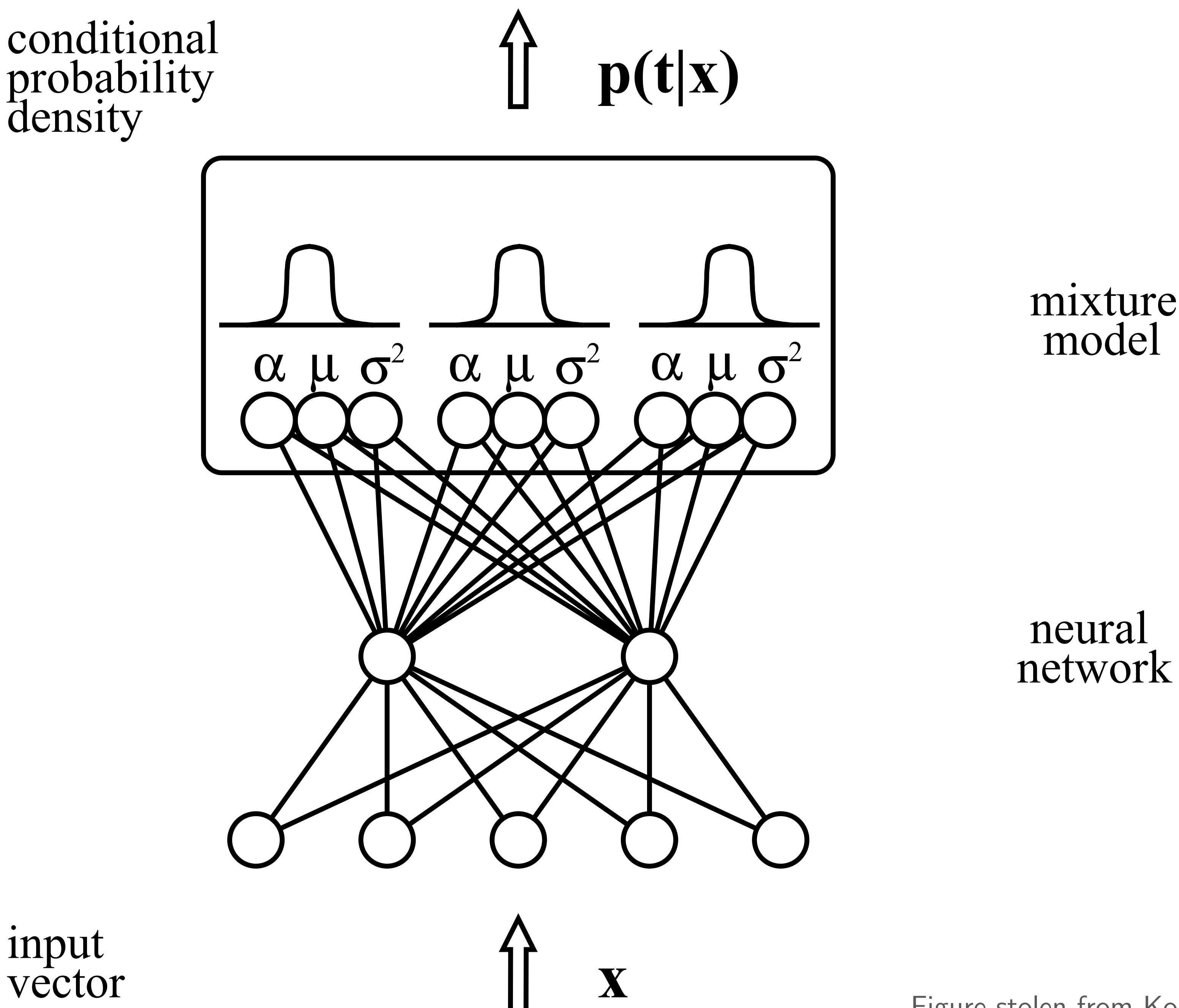


Figure stolen from Korin Richmond

Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation

Papamakarios and Murray (NIPS, 2016)
Lueckmann et al. (NIPS, 2017)

— Fit $\hat{p}(\theta \,|\, \mathbf{x})$ maximize $\sum_s \log \hat{p}(\theta^{(s)} \,|\, \mathbf{x}^{(s)})$

— $\hat{p}(\theta \,|\, \mathbf{x}_{\text{observed}}) \;\rightarrow\;$ approx posterior

Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation

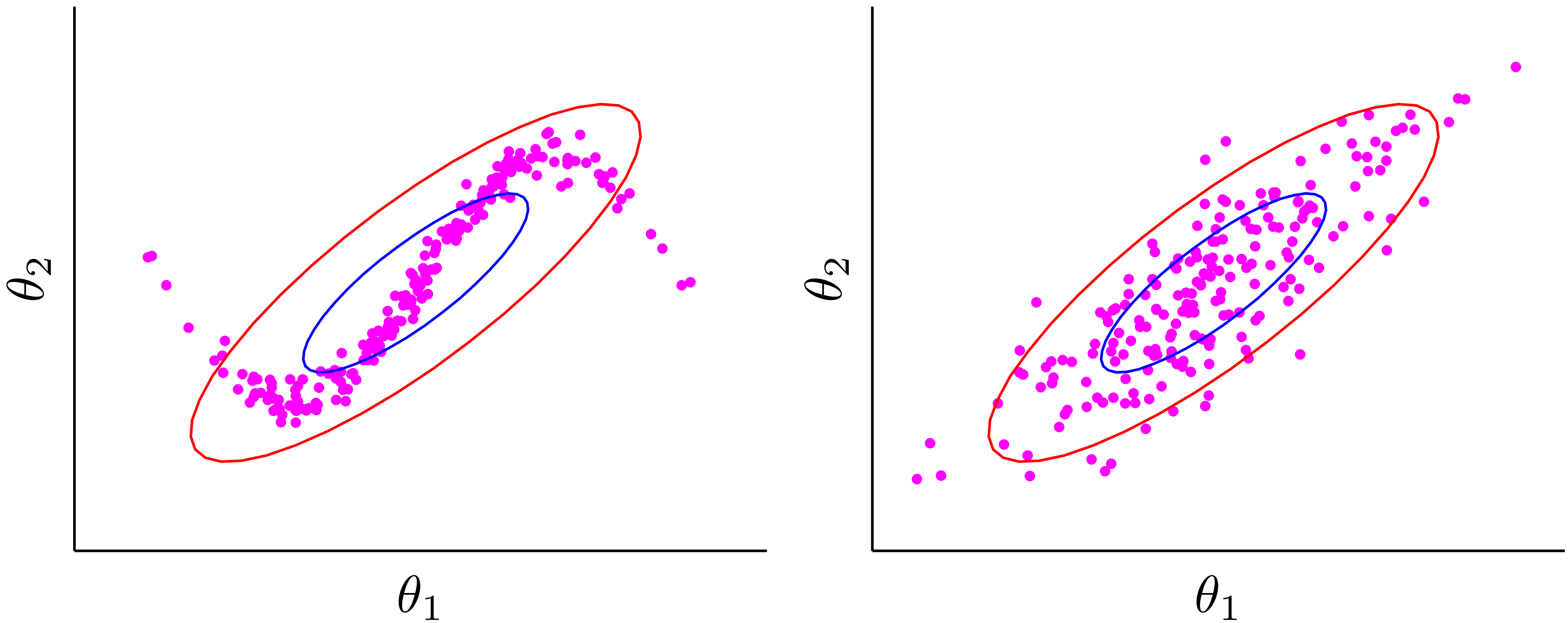
Papamakarios and Murray (NIPS, 2016)
Lueckmann et al. (NIPS, 2017)

— Fit $\hat{p}(\theta \,|\, \mathbf{x})$ maximize $\sum_s \log \hat{p}(\theta^{(s)} \,|\, \mathbf{x}^{(s)})$

— $\hat{p}(\theta \,|\, \mathbf{x}_{\text{observed}}) \;\rightarrow\;$ approx posterior

— Refine fit: more simulations

Underfitting



True posterior samples

samples from Gaussian fit

— Modeling posteriors

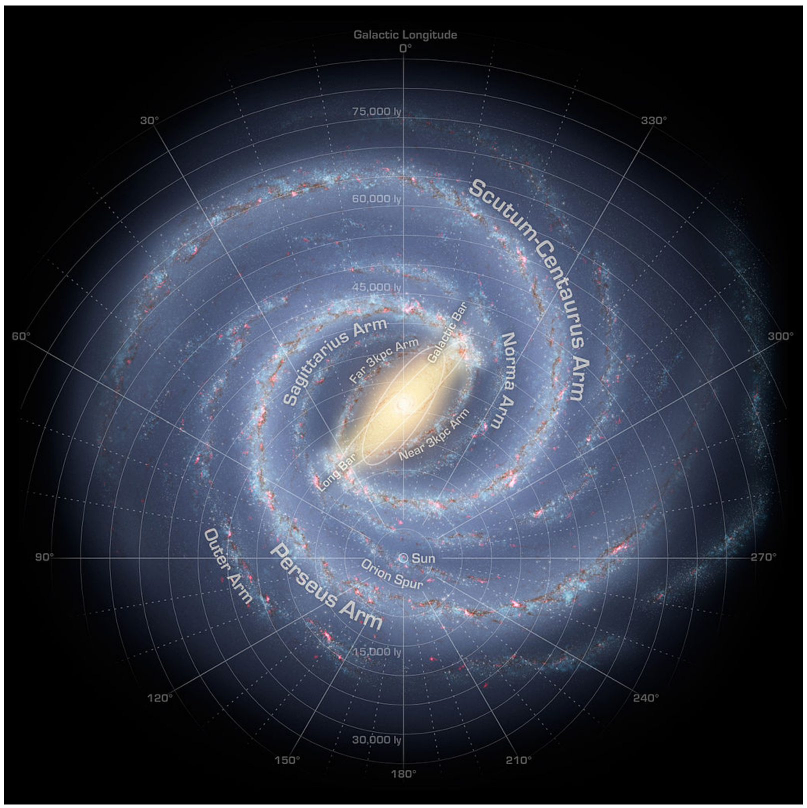
— **Modeling priors**

— Modeling likelihoods

Weighing the Milky Way

Busha, Marshall, Wechsler, Klypin and Primack (2011)

APJ 743:40



Milky Way diagram, NASA

Magellanic Clouds, ESO/S. Brunier

http://en.wikipedia.org/wiki/File:236084main_MilkyWay-full-annotated.jpg
http://www.eso.org/public/images/b01/

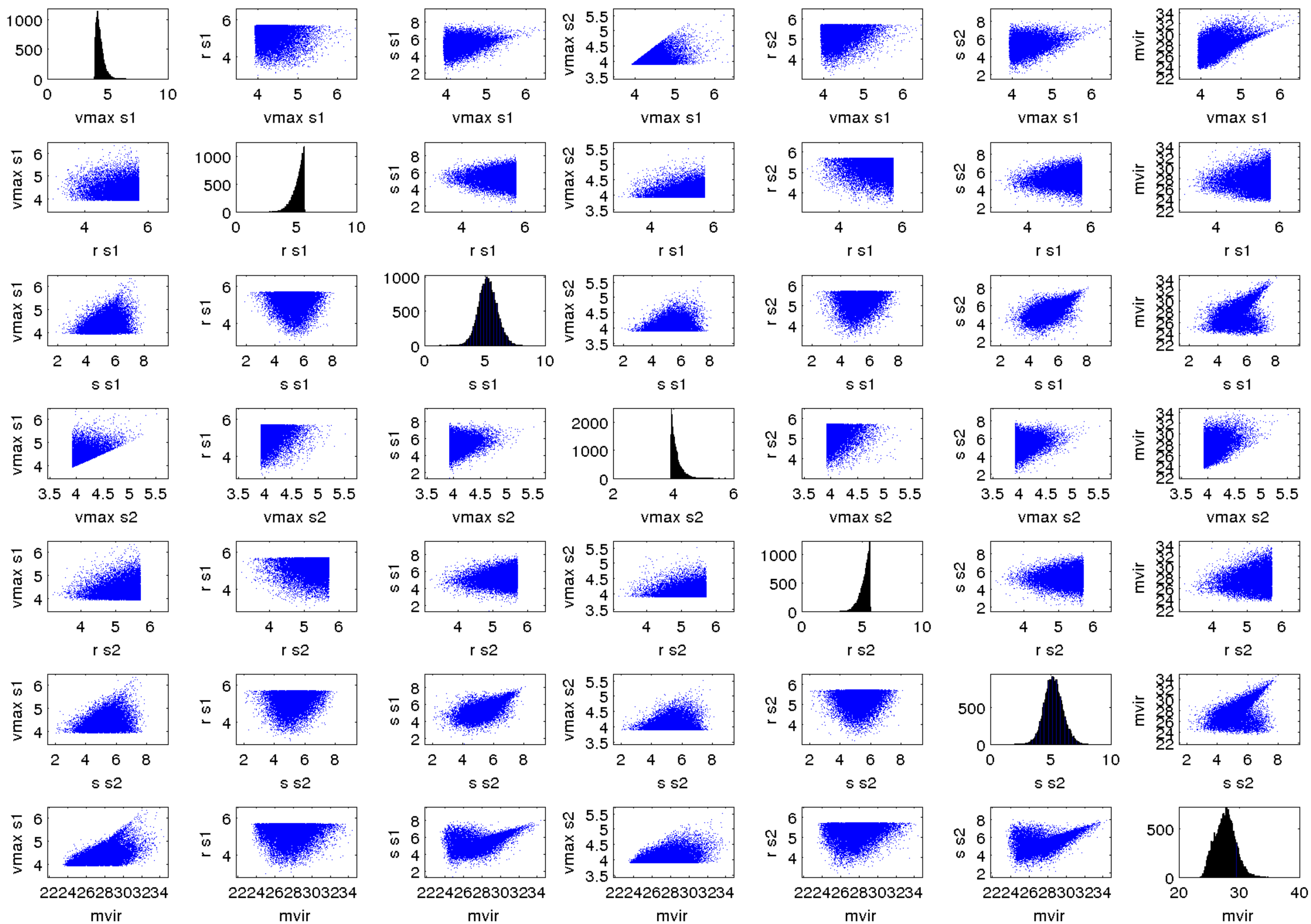
Bayesian Inference

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) \, p(\mathbf{x})$$

$\mathbf{x} = [r, v, m]$, vector of galaxy properties

$\mathbf{y} = [\hat{r}, \hat{v}]$, noisily observe part of $\mathbf{x}$

The prior: simulation samples



Bayesian inference

What is our Galaxy like?

1. Sample from prior

Imaginary galaxies with mass and companion galaxies

2. Weight samples with likelihood

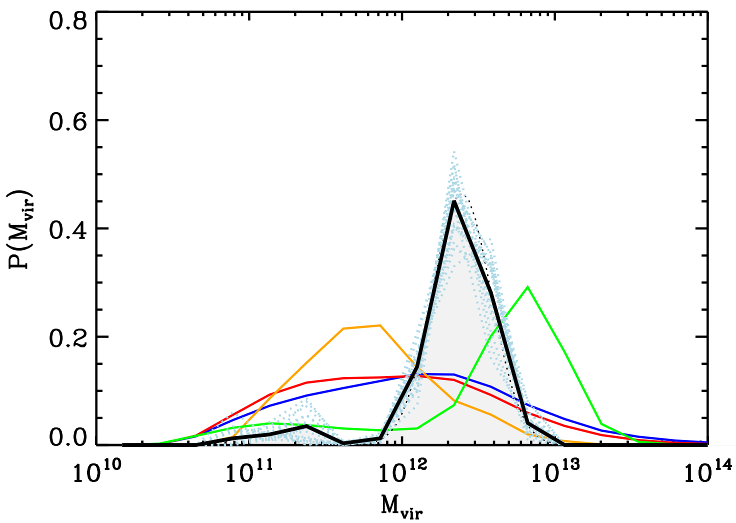
Chuck out galaxies without companions like ours

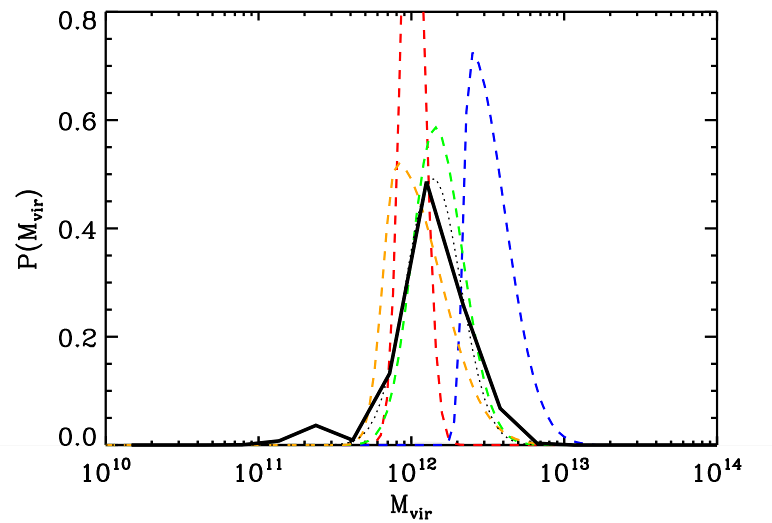
3. Use weighted samples

Look at masses of remaining galaxies

That is, do simple importance sampling

Existing answer





2.1 million simulated galaxies

36,000 with two companions

400 within 2σ of Milky Way observations

Busha et al., arXiv:1011.2203v3

Simulations are data. . .

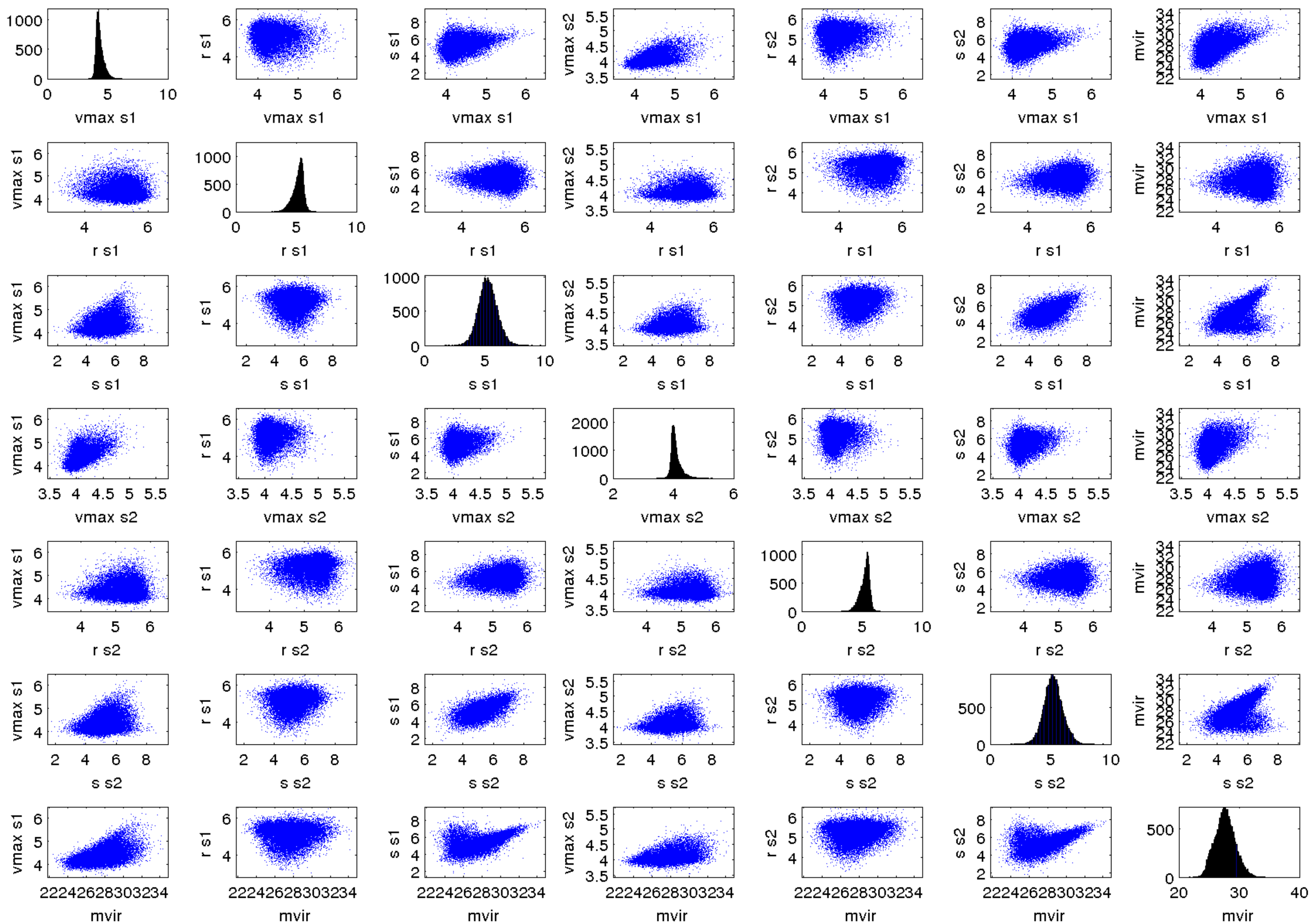
$$P(\mathbf{x} \,|\, \mathbf{y}, \mathcal{S}) \propto P(\mathbf{y} \,|\, \mathbf{x})\, P(\mathbf{x} \,|\, \mathcal{S})$$

$\mathbf{x} = [r, v, m]$, vector of galaxy properties

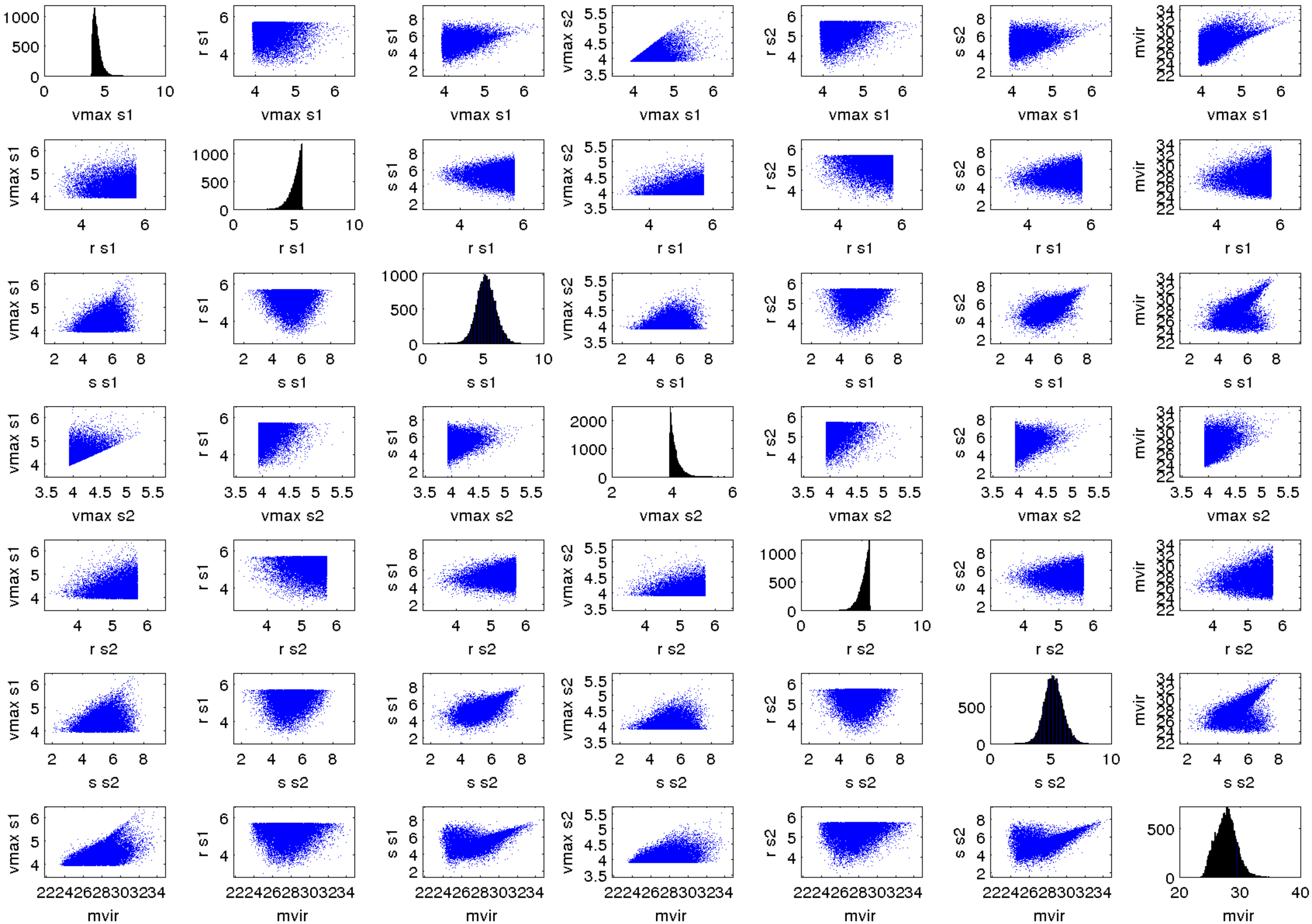
$\mathbf{y} = [\hat{r}, \hat{v}]$, noisily observe part of $\mathbf{x}$

$\mathcal{S} = \{\mathbf{x}^{(s)}\}$, simulated galaxy vectors

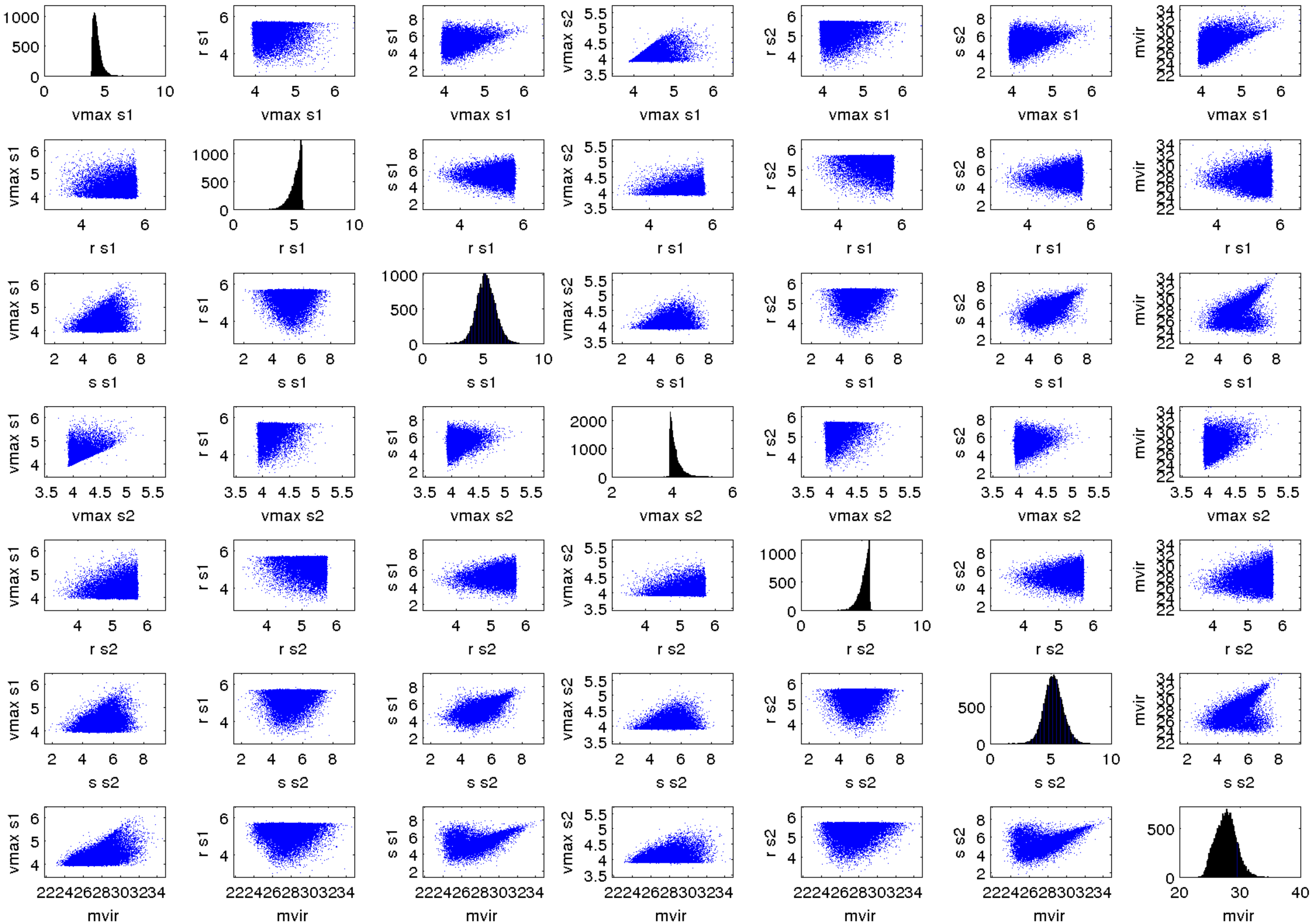
Mixture of Gaussian samples



Simulation samples



AMDN samples

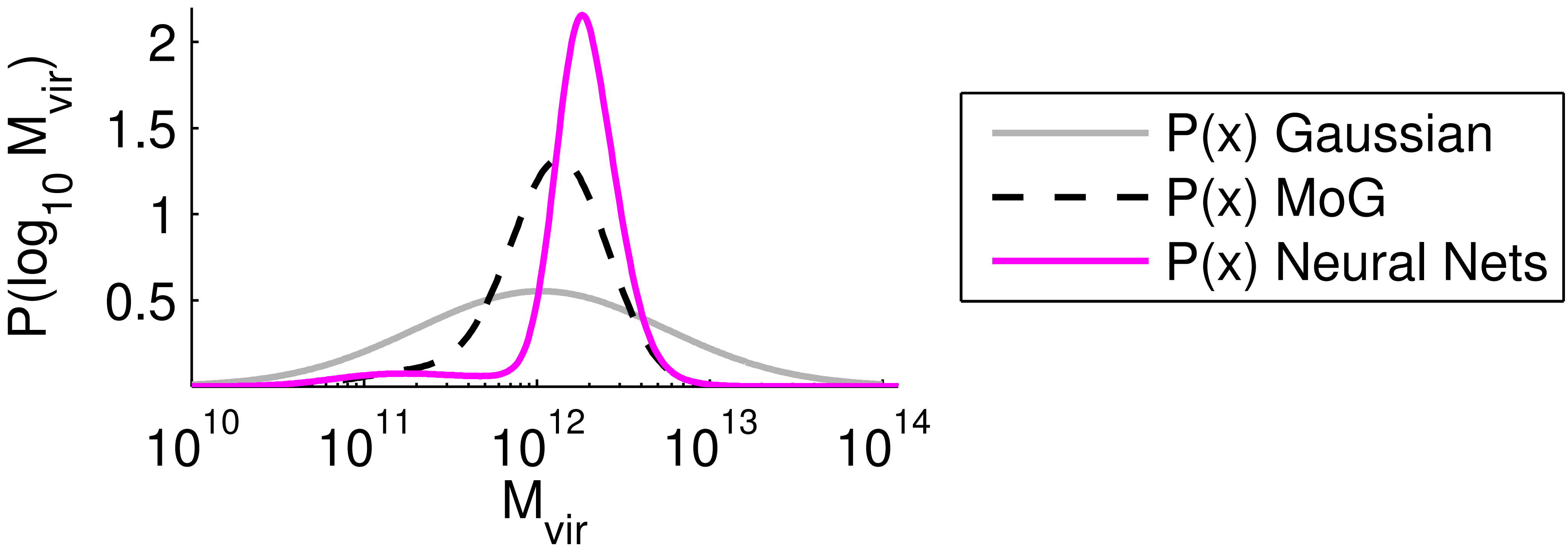


Milky Way mass

$p(\mathbf{x})$ theory: simulated galaxy properties

$p(\mathbf{y} \mid \mathbf{x})$ observations of Milky Way

$p(\mathbf{x} \mid \mathbf{y}) \rightarrow p(x_1 \mid \mathbf{y})$, posterior of mass

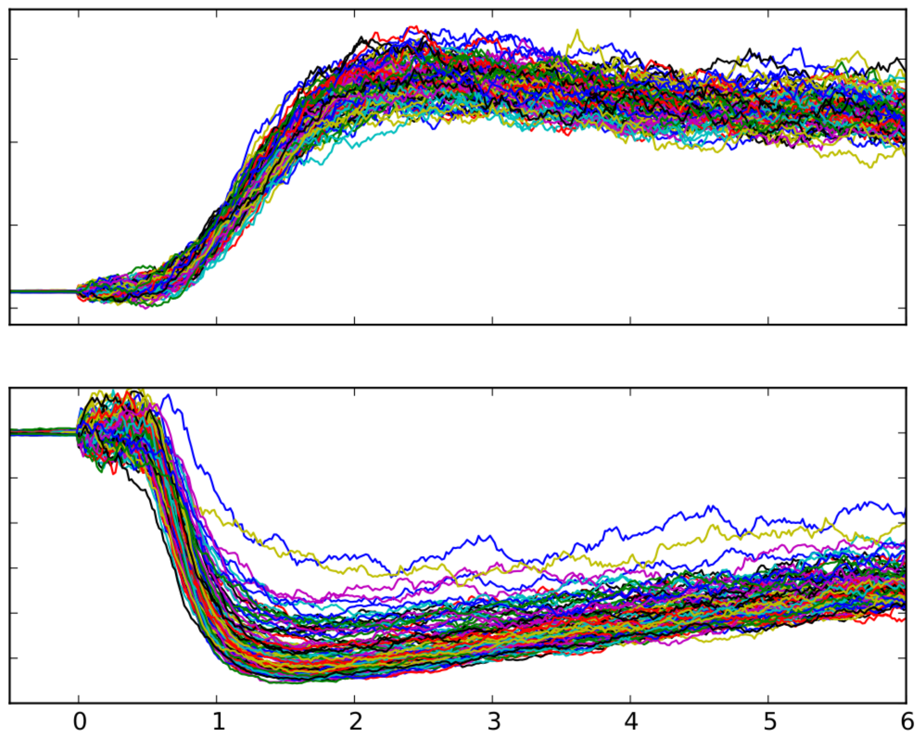


— Modeling posteriors

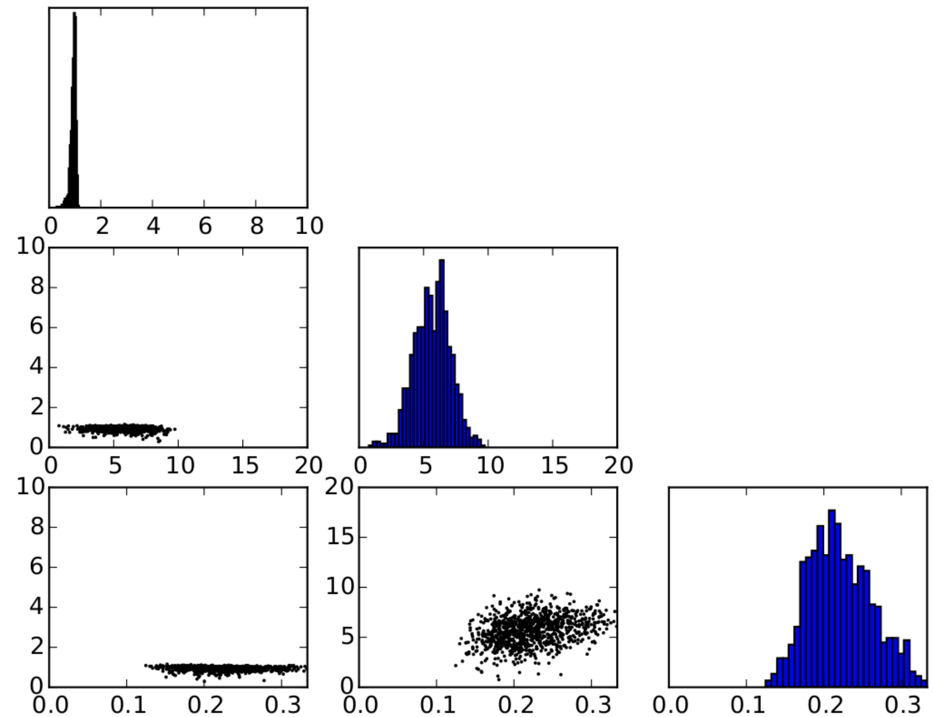
— Modeling priors

— **Modeling likelihoods**

Surrogate modeling / emulation



data $\mathbf{x}$



parameters $\boldsymbol{\theta}$

$$p(\theta \,|\, \mathcal{D}) \;\propto\; p(\theta) \prod_n p(\mathbf{x}^{(n)} \,|\, \theta)$$

Cf Cranmer, Pavez, Louppe (2016) arXiv:1506.02169

Thanks!

`http://iainmurray.net`

NADE variants, MADE, and MAF

ϵ-free ABC, pseudo-marginal slice sampling

Can do ABC by density estimation

... or conditional density estimation

Neural Autoregressive Models can do both

— Larochelle and Murray (2011)

— Uría, Murray, and Larochelle (2013, 2014)

— Germain, Gregor, Murray, Larochelle (2015)

Building autoregressive models

Lots of credit due elsewhere:

. . .

Frey et al. (NIPS 1996), Frey (book, 1998)

Bengio and Bengio (NIPS, 2000)

Li and Stephens (Genetics, 2003)

. . .

Modeling via the Chain Rule

$p(x_1)$

$p(x_2 \mid x_1)$

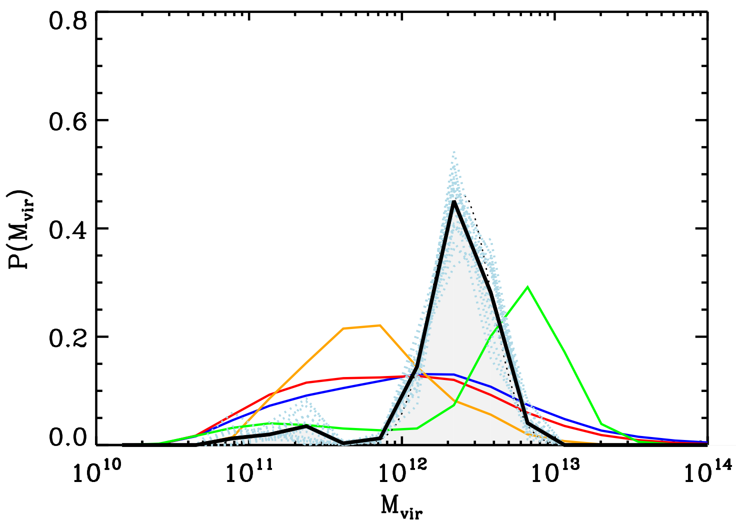
$p(x_3 \mid x_1, x_2)$

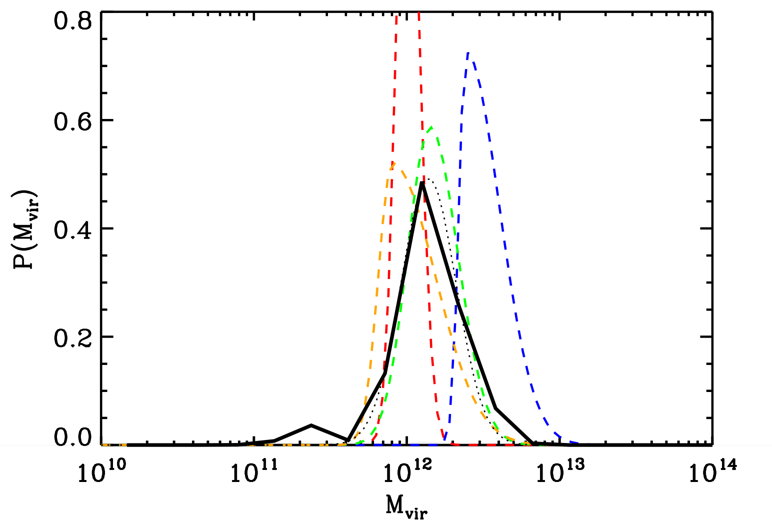
$p(x_4 \mid x_1, x_2, x_3)$

$$p(\mathbf{x}) = p(x_1) \prod_{d=2}^{D} p(x_d \mid \mathbf{x}_{<d})$$

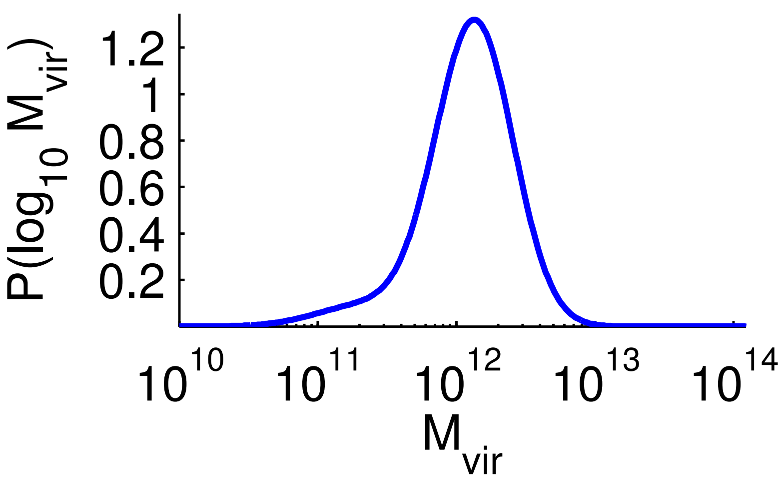
(conditional version straightforward)

Results of inference

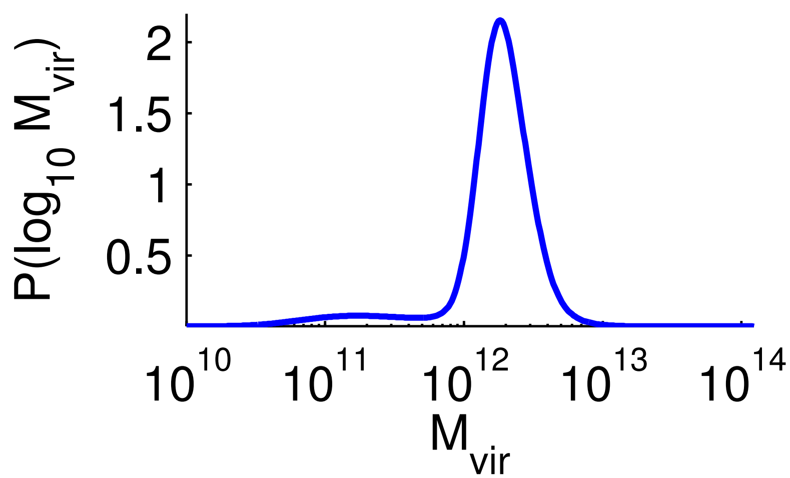




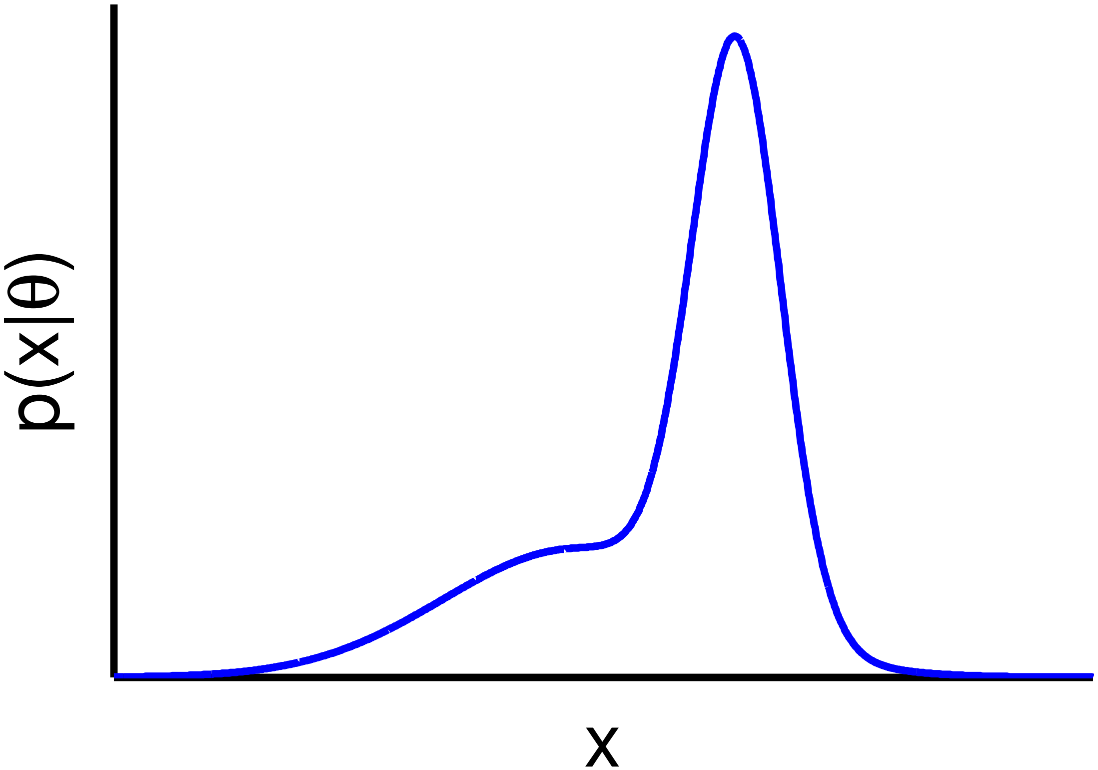
Prior modeled with MoG

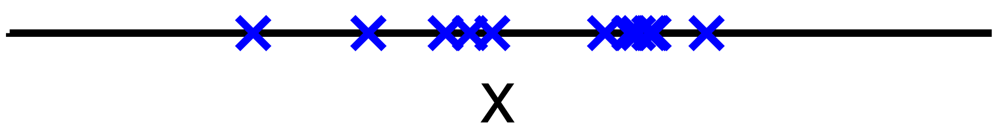


Prior modeled with AMDN

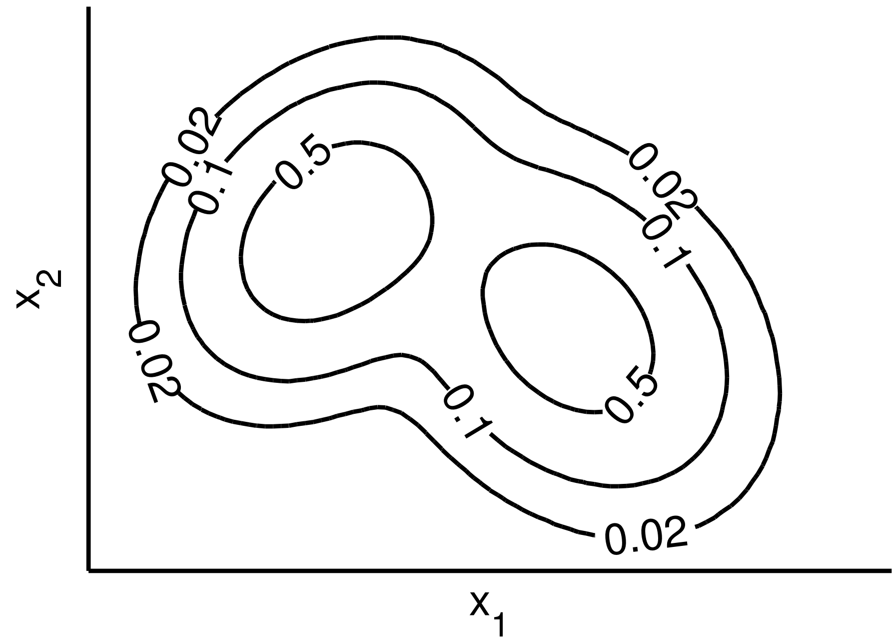


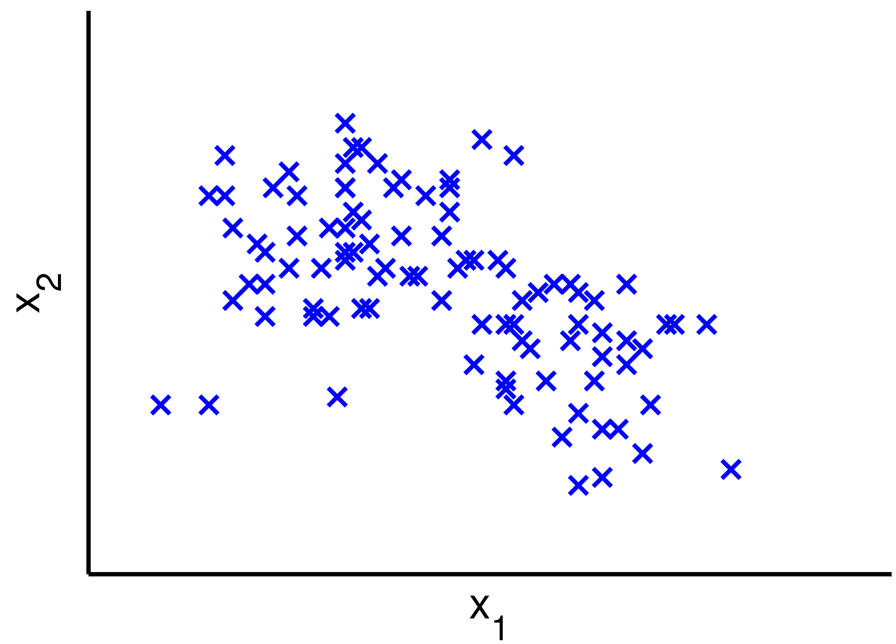
Density estimation





Quantiles of $p(\mathbf{x} \,|\, \theta)$





Task: $\{\mathbf{x}^{(n)}\} \rightarrow \theta$

Conditional density estimation

Can simulate:

$\Omega \;\rightarrow\;$ Universe $\;\rightarrow\; \mathcal{D}$, photons in CCD

Want: $\quad p(\Omega \,|\, \mathcal{D})$

Application to weak lensing

Can simulate:

$$\Omega \;\rightarrow\; \text{Universe} \;\rightarrow\; \text{photons in CCD, } \mathcal{D}$$

$$\downarrow$$

$$\text{Shear statistics, } \hat{\xi}$$

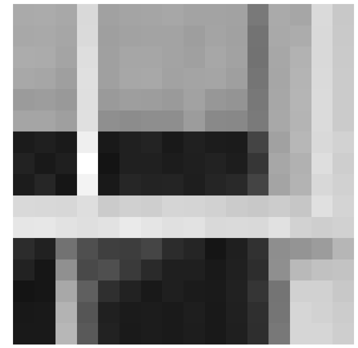
Learn: $\quad p(\Omega \,|\, \hat{\xi})$

$= p(\Omega \,|\, \mathcal{D})$ if $\hat{\xi}$ a 'sufficient statistic'

Cf *Approximate Bayesian Computation via regression density estimation*, Fan et al., Stat 2013. Also much older *'recognition networks'*.

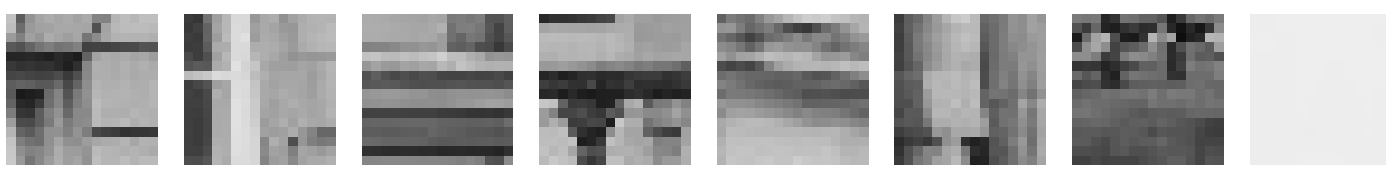
Example: Image denoising

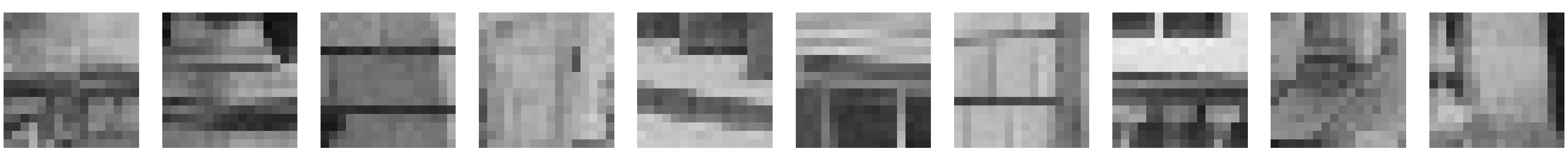
$\mathbf{y}$: 

$\mathbf{x}$: 

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$$

Likelihood: e.g. $\mathcal{N}(\mathbf{y};\, \mathbf{x},\, \sigma^2 I)$

Prior samples: 

 $\ldots$

Zoran and Weiss, ICCV 2011

(a) Blurred (b) Krishnan et al. (c) EPLL GMM

$p(\mathbf{x})$ = Mixture of Gaussians fitted to patches

The likelihood: observations

Table 1
Observed Properties of the LMC and SMC

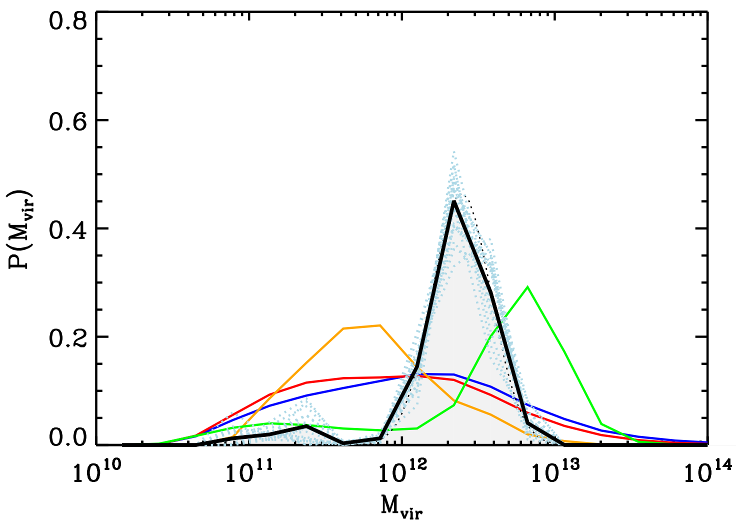
Property	LMC	SMC	Reference
$v_{\rm max}$ (km s^{-1})	65 ± 15	60 ± 15	vdM02, S04, HZ06
r_0 (kpc)	50 ± 2	60 ± 2	vdM02
s (km s^{-1})[a]	378 ± 36	301 ± 104	K06

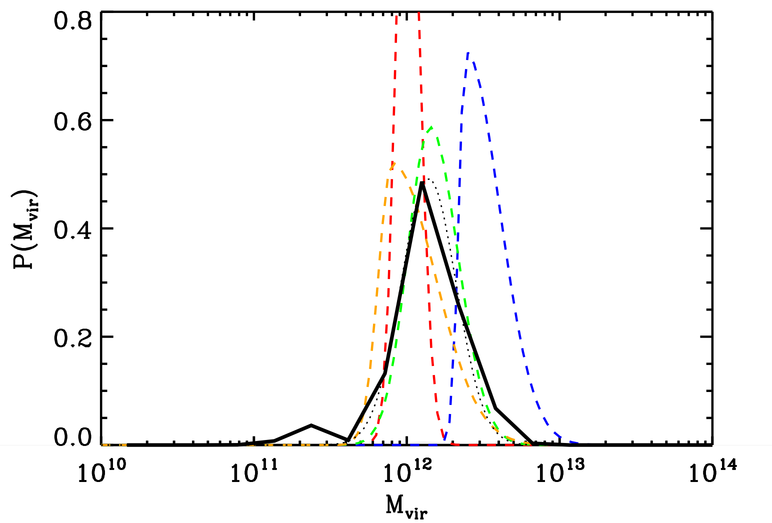
Notes. For a given satellite, $v_{\rm max}$ is its estimated maximum circular velocity, r_0 is its estimated distance from the Galactic center, and s is its estimated speed relative to the Galactic center. References are vdM02 = van der Marel et al. (2002); S04 = Stanimirović et al. (2004); K06 = Kallivayalil et al. (2006a, 2006b); HZ06 = Harris & Zaritsky (2006).

[a] Errors on s have been increased relative to the published values for conservatism (see the text).

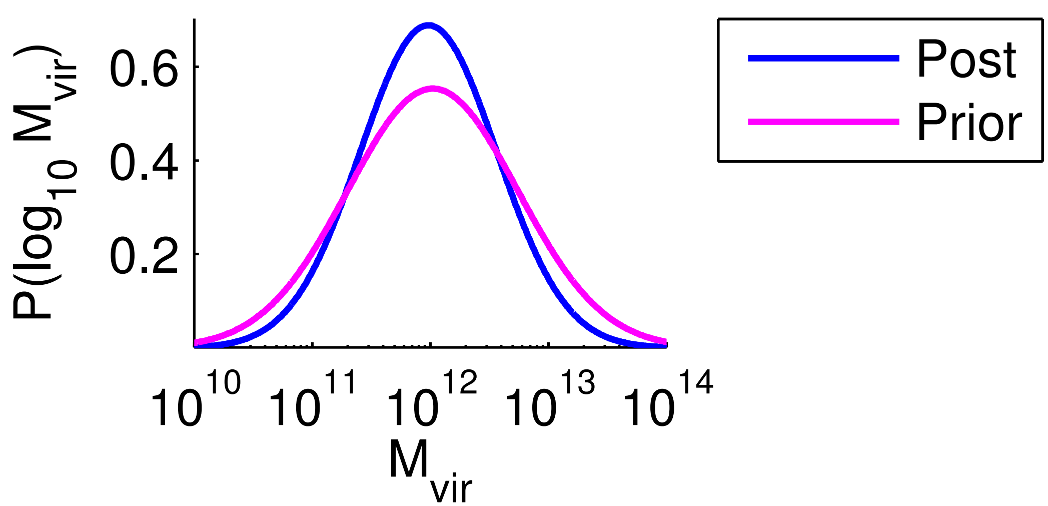
— Busha et al. (2011), APJ 743:40

Parametric assumptions

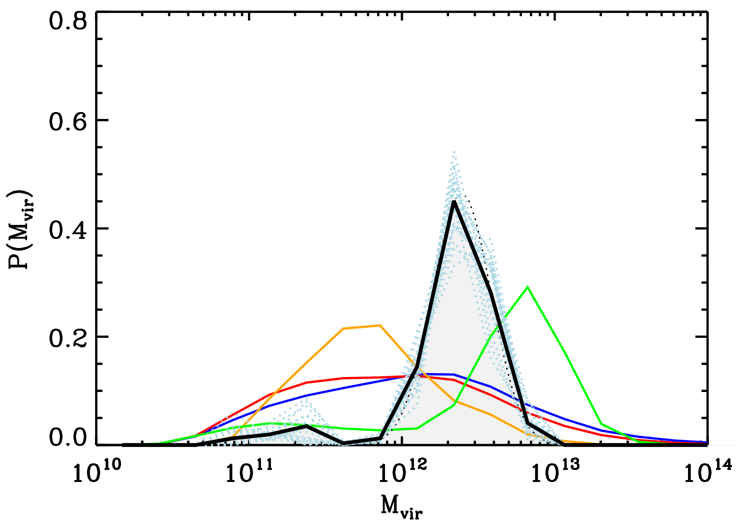


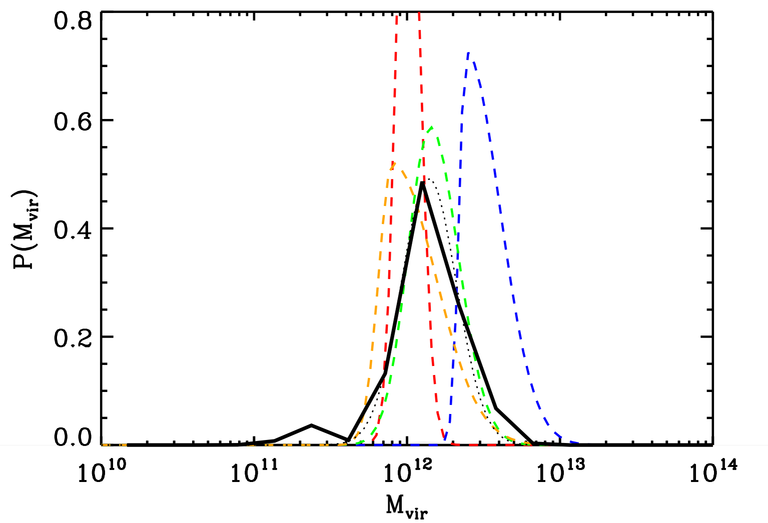


Assume prior is a Gaussian

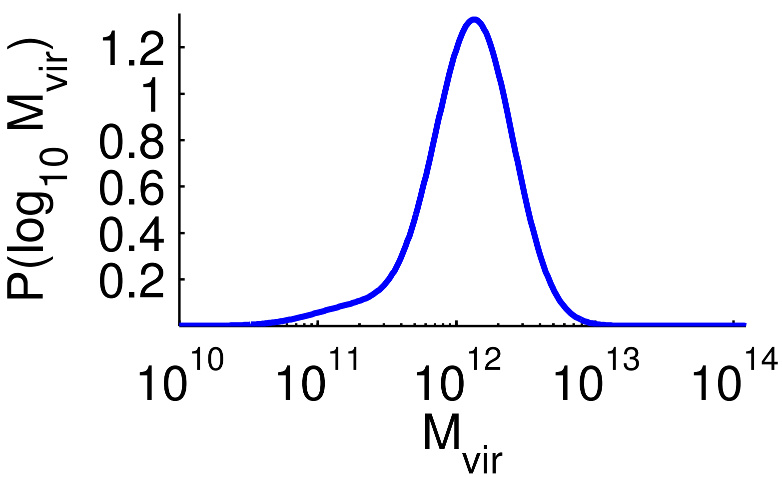


Parametric assumptions

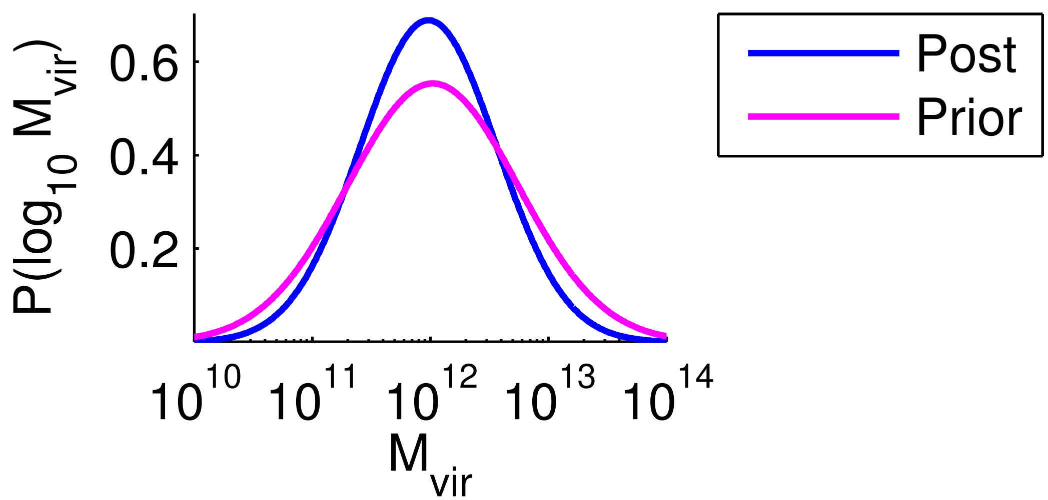




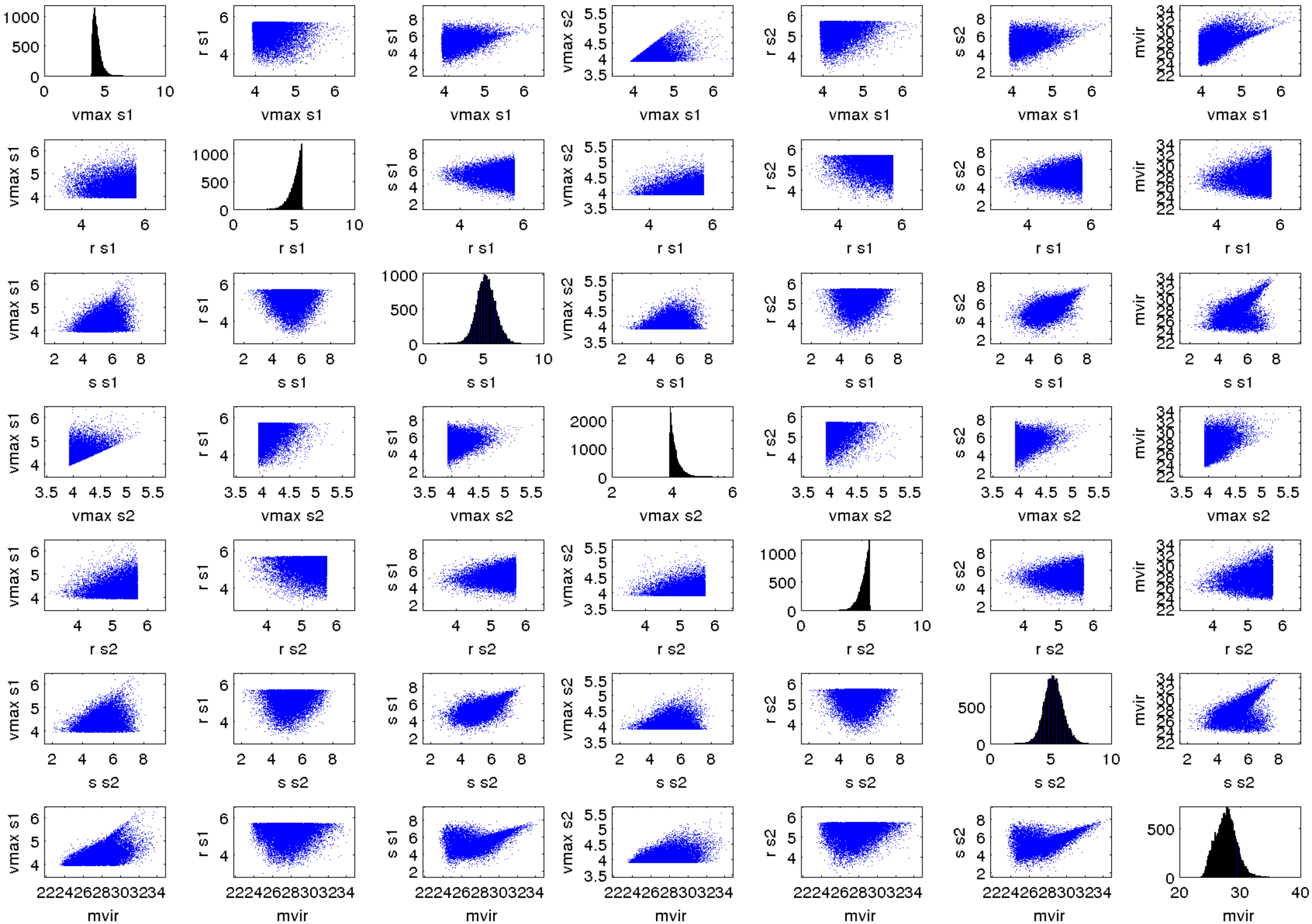
Prior modeled with MoG



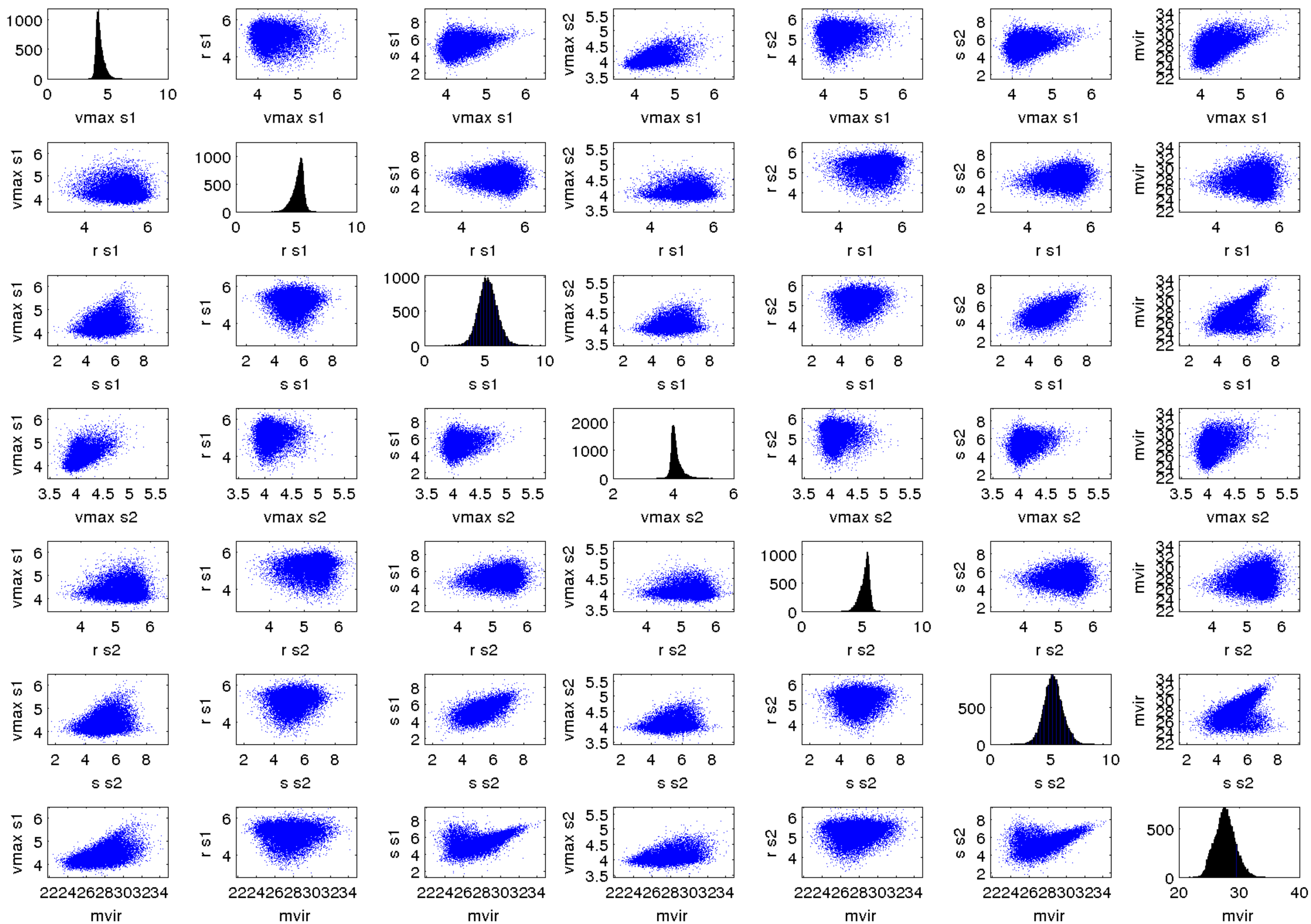
Assume prior is a Gaussian



Simulation samples



Mixture of Gaussian samples



Disclaimer

I *like* mixtures of Gaussians!

Zoran & Weiss ICCV 2011 — denoising/deblurring images

Bovy, Hogg, Roweis 2011 — Extreme deconvolution

Hogg & Lang, 2013 — Replacing Standard Galaxy Profiles with Mixtures of Gaussians

. . .

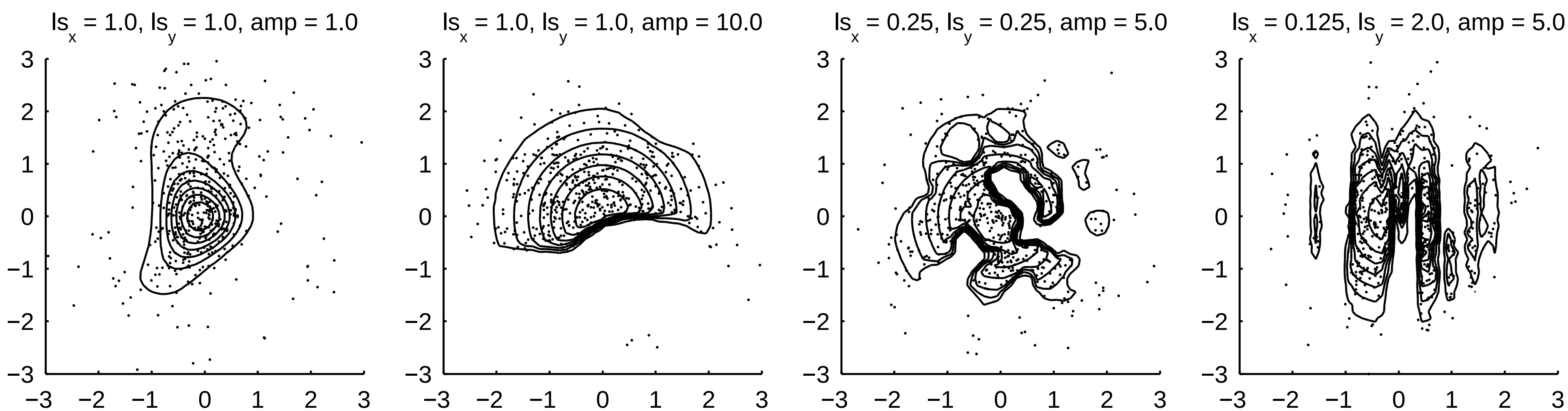
GP Density estimation

$$p(x|\mathbf{f}) = \frac{1}{\mathcal{Z}(\mathbf{f})}\Phi(\mathbf{f}(x))\pi(x)$$

$\mathbf{f} \sim \mathcal{GP}$

$\Phi =$ sigmoidal function

$\pi =$ base measure



Gaussian Process Density Sampler

Adams, Murray and MacKay (2009).