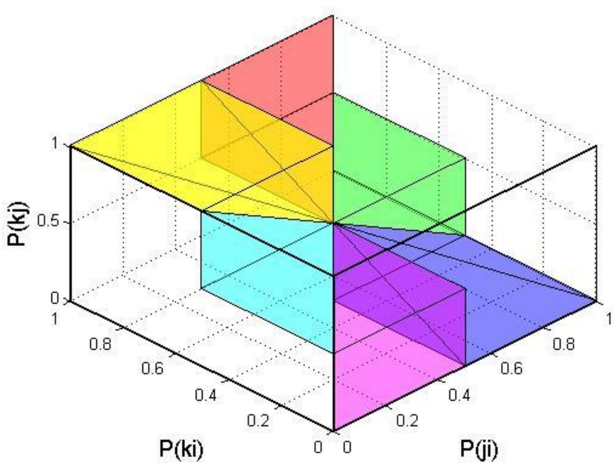
Probabilistic Specification and Quantitative Testing of Decision Theories: *Bayesian Approaches*

Daniel R. Cavagnaro
California State University, Fullerton



Presented at the 2016 Meeting of the European Mathematical Psychology Group
June 19, 2016; Copenhagen, Denmark

Why Bayesian?

- The Bayesian framework provides a unified approach to addressing three important issues in statistical inference:
  - Model Estimation
    - Which parameters best fit the data?
  - Model Evaluation
    - Is the fit adequate?
  - Model Selection
    - Which model is best (at approximating the underlying mental process)?

Outline

- *Bayesian* Probabilistic Specification of a (Binary Choice) Decision Theory
- *Bayesian* Quantitative testing of decision theories
  - Model Estimation, Evaluation, and Selection
- Bayesian methods for evaluating group data

Bayesian Probabilistic Specification of a (Binary Choice) Decision Theory

- Notation and Assumptions:
  - $\mathcal{D}$ = a domain of d distinct unordered pairs of choice options.
  - N_{xy} = the number of times that the pair $\{x, y\}$ is presented to the decision maker
  - n_{xy} = the number of times that x was chosen from $\{x, y\}$
  - $\boldsymbol{n} = \{n_{xy}\}_{\{x,y\}\epsilon\mathcal{D}}$

Bayesian Probabilistic Specification of a (Binary Choice) Decision Theory

- Notation and Assumptions:
  - P_{xy} = the binary choice probability of x being chosen from $\{x, y\} \epsilon \mathcal{D}$.
  - $\boldsymbol{P} = \{P_{xy}\}_{\{x,y\} \epsilon \mathcal{D}}$
  - Repeated choices on the same choice pair are *exchangeable*
  - Choices on distinct choice pairs are independent

Bayesian Probabilistic Specification of a (Binary Choice) Decision Theory

- Likelihood function:

$$f(\boldsymbol{P}|\boldsymbol{n}) = \prod_{x,y\epsilon\mathcal{C}} \binom{N_{xy}}{n_{xy}} P_{xy}^{n_{xy}} (1 - P_{xy})^{1-n_{xy}}$$

What about the decision theory?

Bayesian Probabilistic Specification of a (Binary Choice) Decision Theory

- Decision theories are distinguished by the *constraints* they place on choice probabilities (i.e., model parameters).

- Any model m restricts $\boldsymbol{P}$ to some parameter subspace $\lambda_m \subseteq [0,1]^d$

- Prior distribution

$$\pi(\boldsymbol{P}|m) = \begin{cases} \frac{1}{v_m} & \text{if} \quad \boldsymbol{P} \epsilon \lambda_m \\ 0 & \text{otherwise} \end{cases}$$

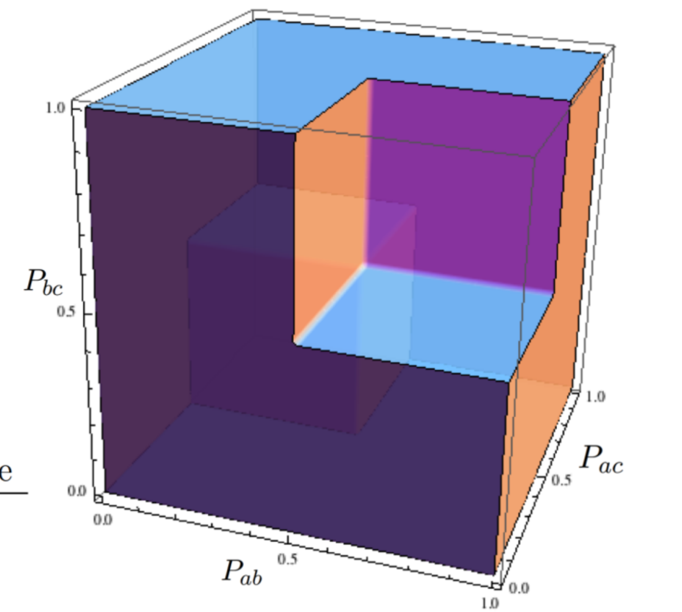
Where v_m is the “volume” of λ_m

Example:

- Weak stochastic transitivity (WST)
  - For any three choice alternatives a, b, c:

if	$P_{ab} > 0.5$
and	$P_{bc} > 0.5$
then	$P_{ac} > 0.5$

Fig. 1a



$\|\mathcal{C}\|$	Volume
3	.75
4	.37
5	.12

$|\mathcal{C}|$ = Number of choice alternatives.

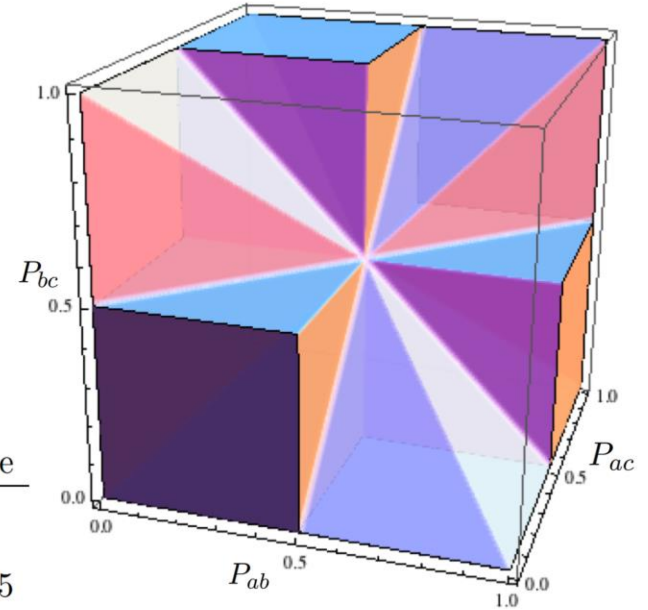
$$|D| = \frac{|\mathcal{C}|(|\mathcal{C}|-1)}{2}$$

Example:

- Strong stochastic transitivity (SST)
  - For any three choice alternatives a, b, c:

if	$P_{ab} > 0.5$
and	$P_{bc} > 0.5$
then	$P_{ac} > \max\{P_{ac}, P_{bc}\}$

Fig. 1c



$\lvert\mathcal{C}\rvert$	Volume
3	.25
4	.008
5	2.4E-05

Example:

- Null Model (m_0)

For any two choice alternatives *a, b*

$$0 \leq P_{ab} \leq 1$$

- Also referred to as the ***encompassing model***
- Provides a common benchmark for against which to compare substantive models

Summary:

- Bayesian model specification consists of a likelihood function and a prior.
- Each model has the same “product of binomials” likelihood function.
- Specific models are distinguished by the subset of the parameter space on which the prior has support.

Outline

- ~~*Bayesian* Probabilistic Specification of a (Binary Choice) Decision Theory~~

- *Bayesian* Quantitative testing of decision theories
  - Model Estimation, Evaluation, and Selection

- Bayesian methods for evaluating group data

Model Estimation

- Goal: Estimate the *posterior distribution* of the model parameters given the observed data.

- Posterior distribution:

$$\pi(\boldsymbol{P}|\boldsymbol{n}, m) = \frac{p(\boldsymbol{n}|\boldsymbol{P})\pi(\boldsymbol{P})}{p(\boldsymbol{n})} = \frac{p(\boldsymbol{n}|\boldsymbol{P})\pi(\boldsymbol{P})}{\int_{\Lambda_m} p(\boldsymbol{n}|\boldsymbol{P})\pi(\boldsymbol{P})\, d\boldsymbol{P}}$$

  - How do we find the maximum (MAP), mean, variance, etc?

Model Estimation

- When analytical solutions are unavailable, Monte Carlo simulation is used to sample from the posterior distribution.

  - Method 1: "Draw-and-Test"
    - Sample from posterior distribution of unconstrained model and keep only those that satisfy the constraints

  - Method 2: Constrained Gibbs sampler
    - Sample directly from the constrained full conditionals
    - See Myung, Karabatsos & Iverson (JMP, 2005) for details

Illustration of "Draw and Test" Method

- Suppose we are interested in the model m_1 defined on two choice pairs $\{x, y\}$ and $\{z, w\}$ by $P_{xy} + P_{zw} < 1$.

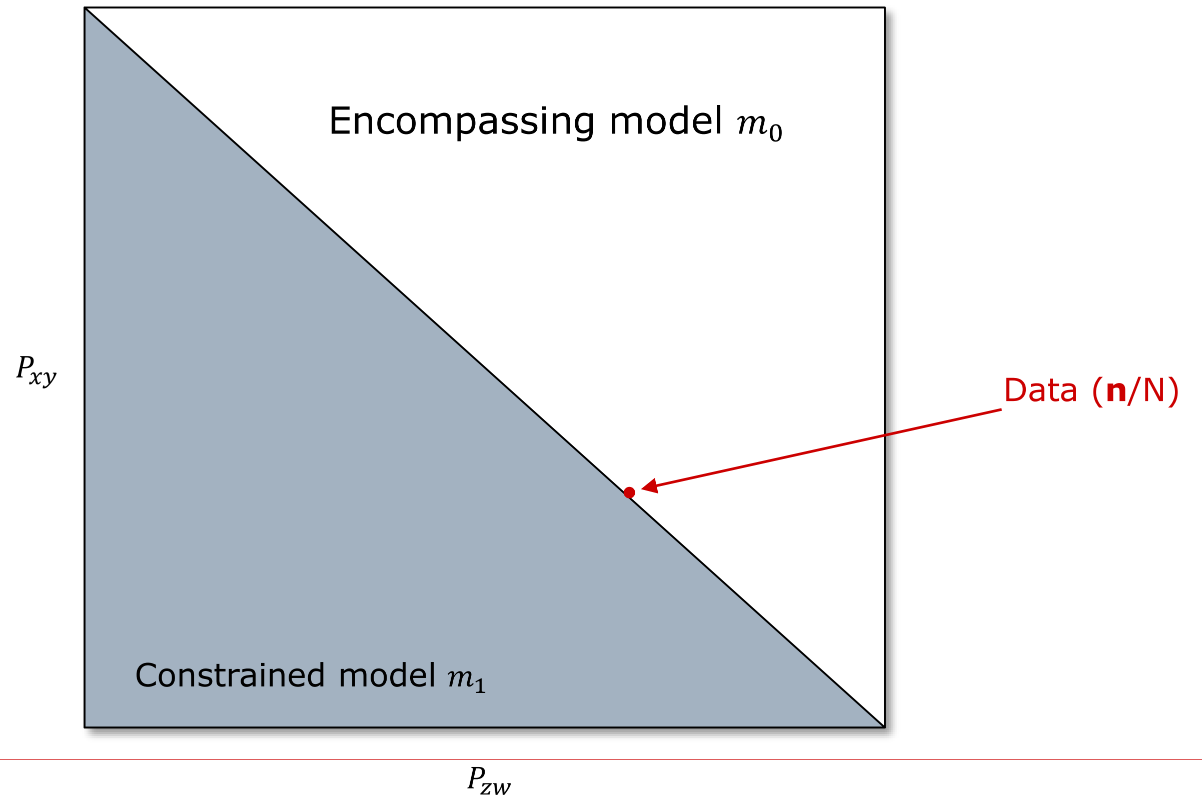


Illustration of "Draw and Test" Method

- Start by drawing a large sample from the posterior of the unconstrained model.

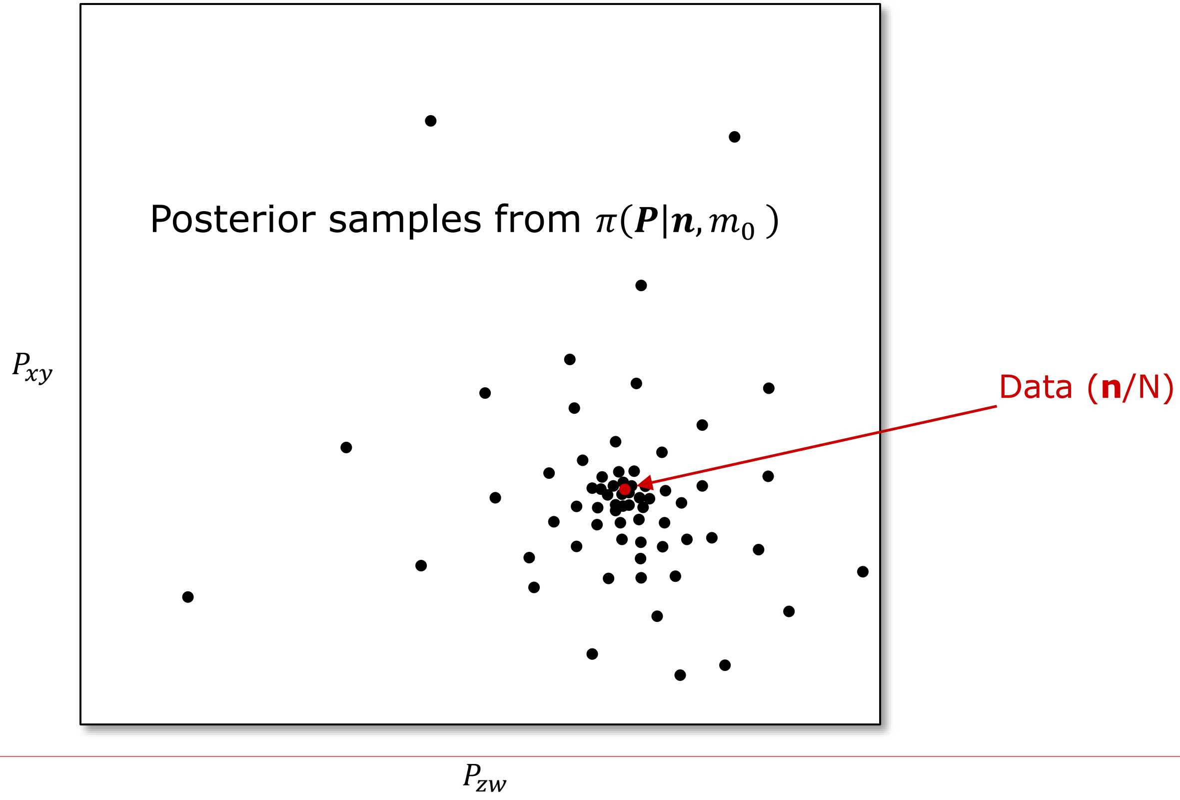


Illustration of "Draw and Test" Method

- Keep only the samples that satisfy $P_{xy} + P_{zw} < 1$.

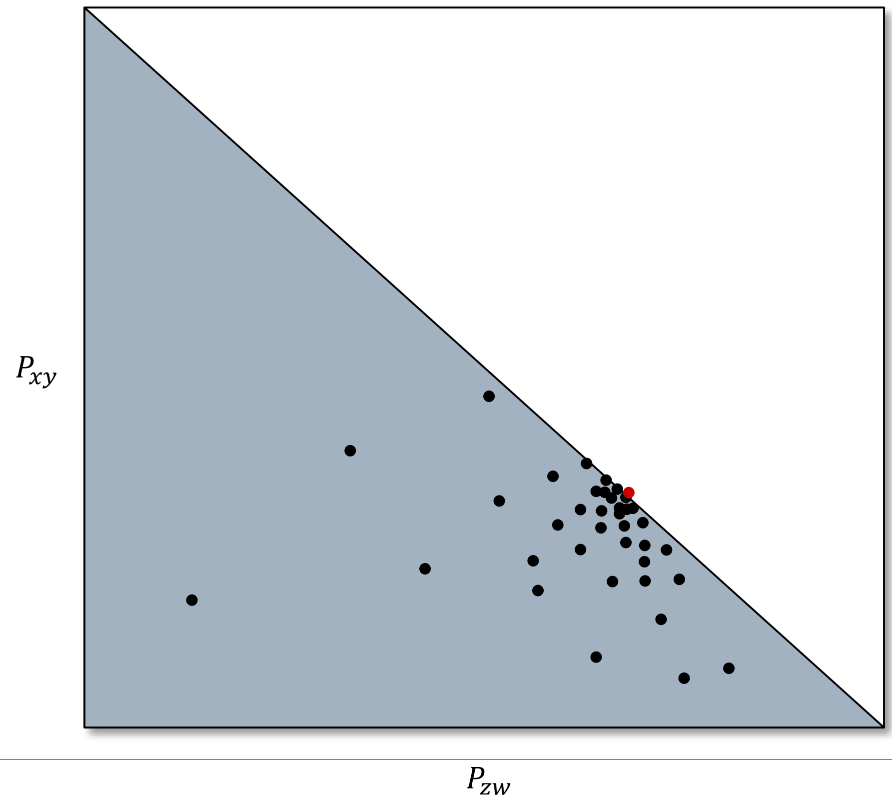


Illustration of constrained Gibbs sampler

- For the constrained Gibbs sampler, begin with any starting value: $\boldsymbol{P}^{(1)} = (P_{zw}^{(1)}, P_{xy}^{(1)})$.

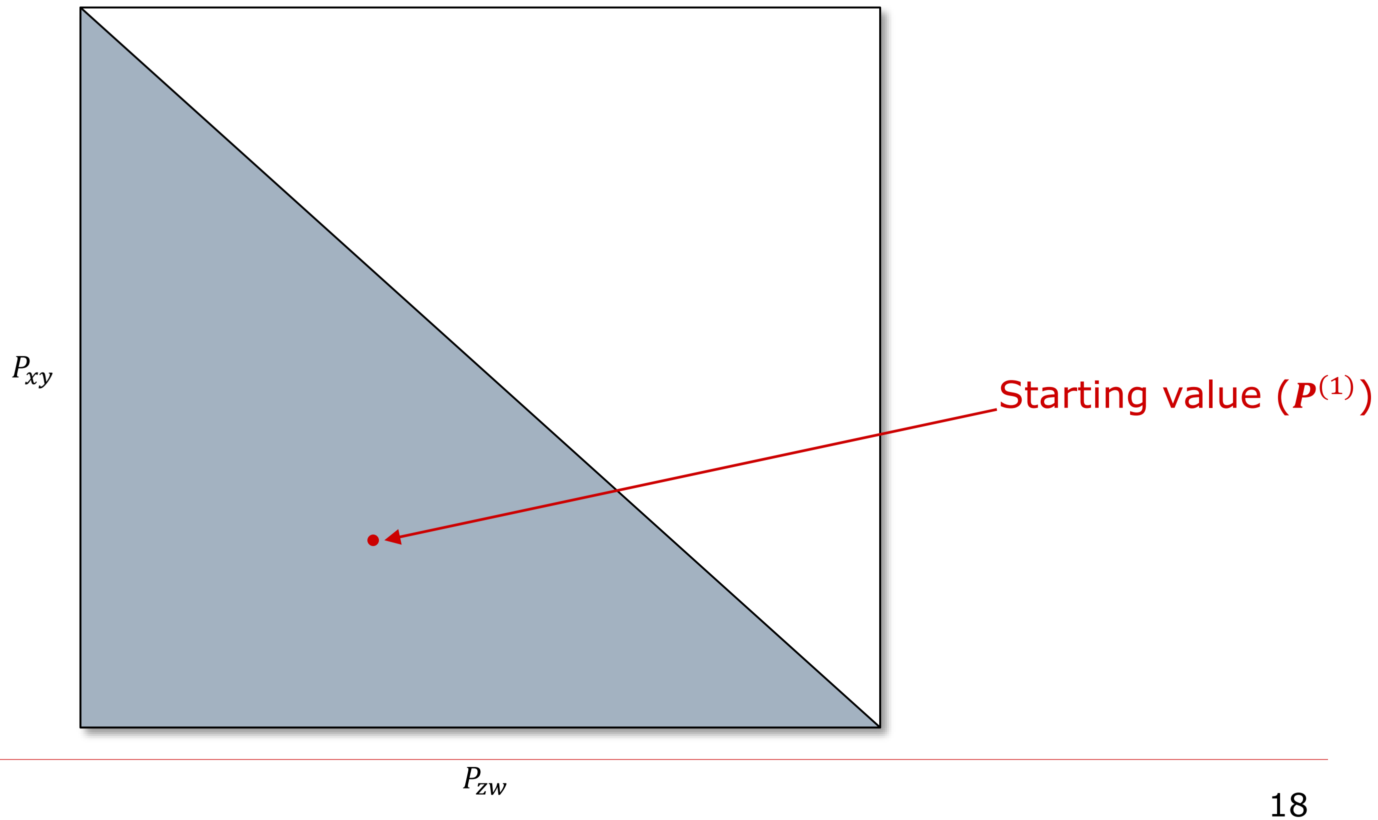


Illustration of constrained Gibbs sampler

- Draw a new value of P_{zw} conditioned on the starting value of P_{xy}.

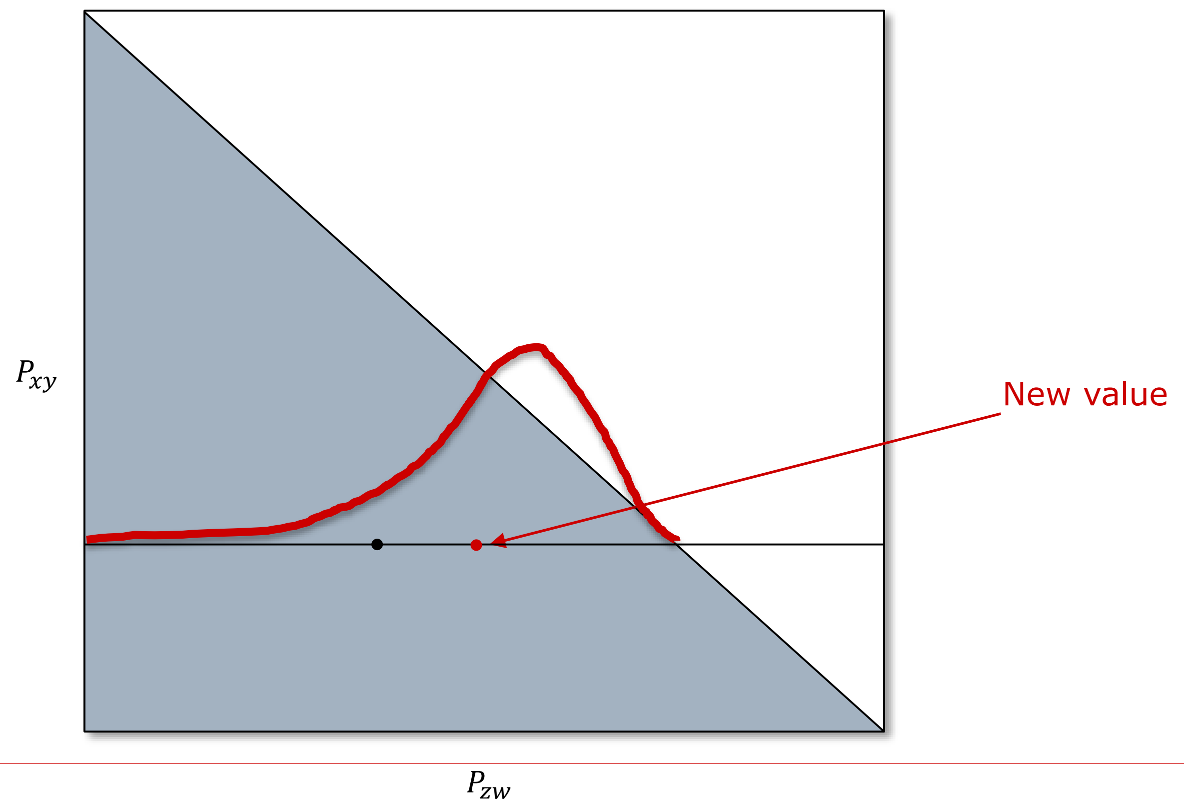


Illustration of constrained Gibbs sampler

- Then, draw a new value of P_{xy} conditioned on the new current value of P_{zw}.

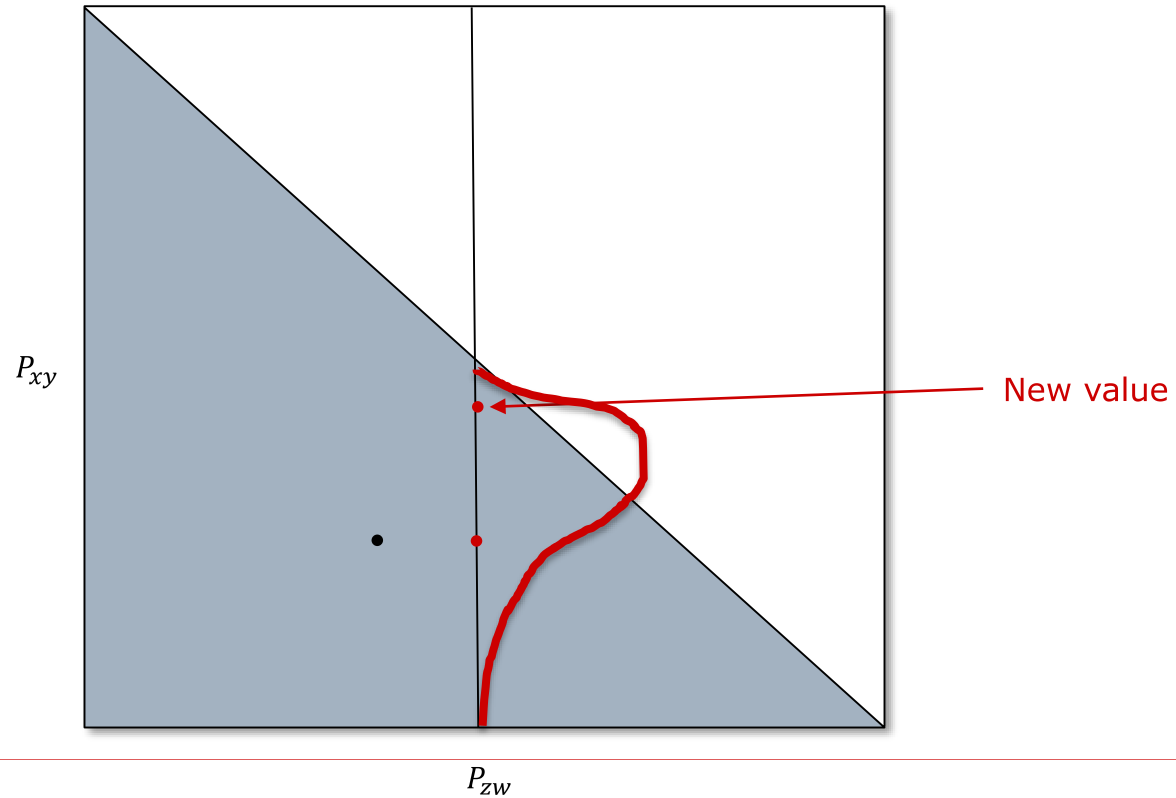


Illustration of constrained Gibbs sampler

- Keep the new pair: $\boldsymbol{P}^{(2)} = (P_{zw}^{(2)}, P_{xy}^{(2)})$.

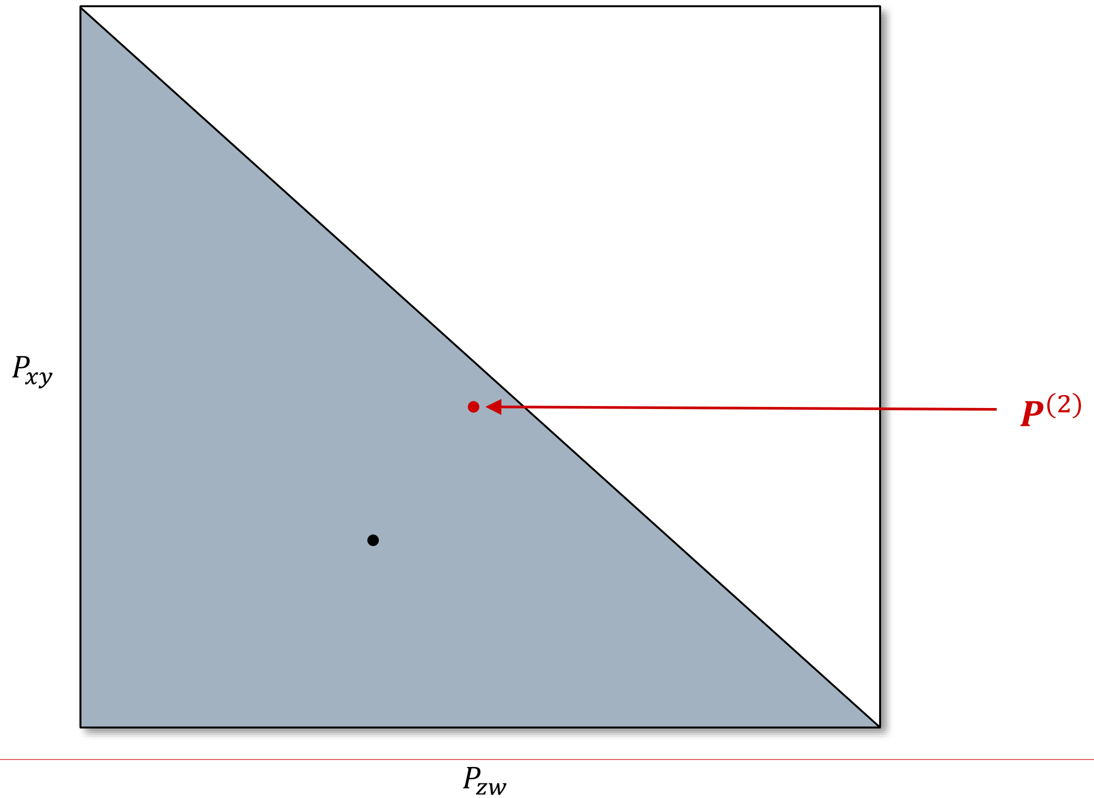


Illustration of constrained Gibbs sampler

- Sample a new value of P_{zw} conditioned on the new current value of P_{xy}.

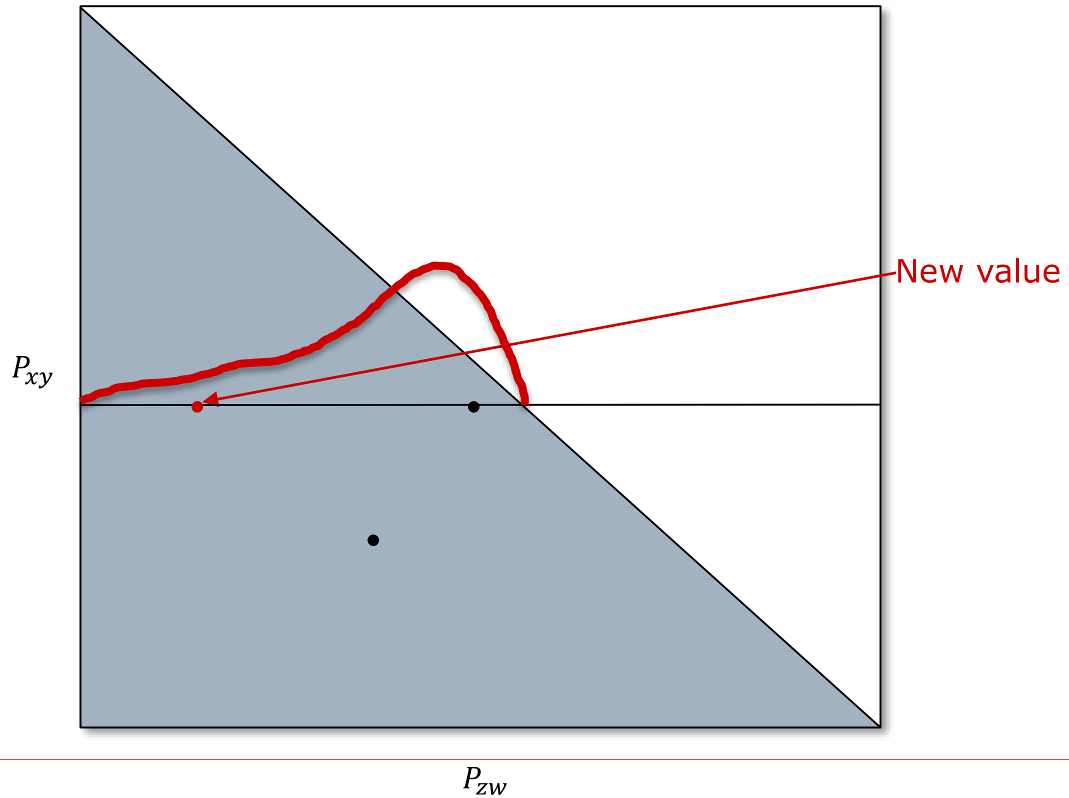


Illustration of constrained Gibbs sampler

- Sample a new value of P_{xy} conditioned on the new current value of P_{zw}.

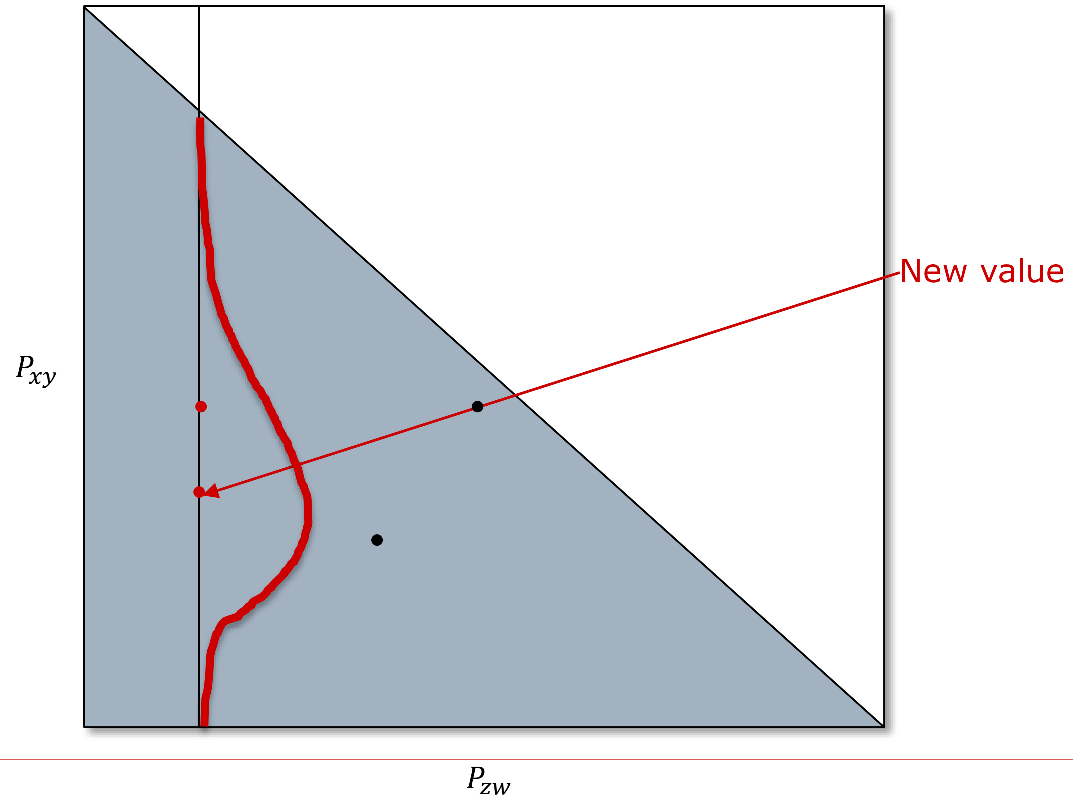


Illustration of constrained Gibbs sampler

- Keep the new pair: $\boldsymbol{P}^{(3)} = (P_{zw}^{(3)}, P_{xy}^{(3)})$.

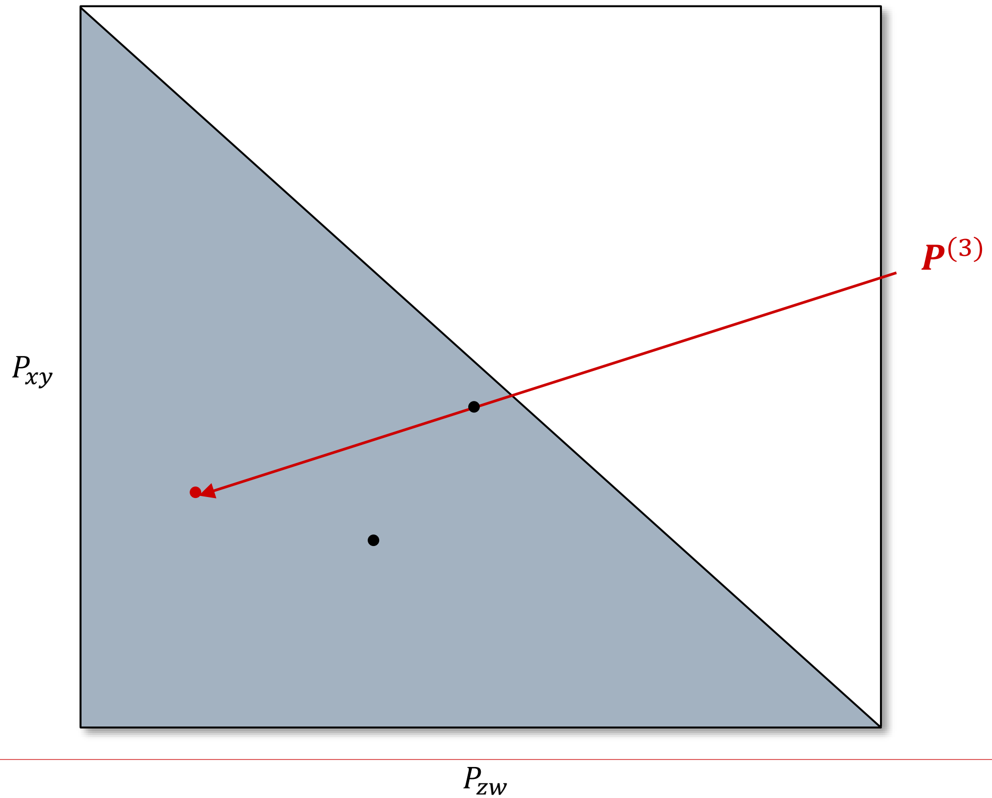


Illustration of constrained Gibbs sampler

- Sample a new value of P_{zw} conditioned on the new current value of P_{xy}.

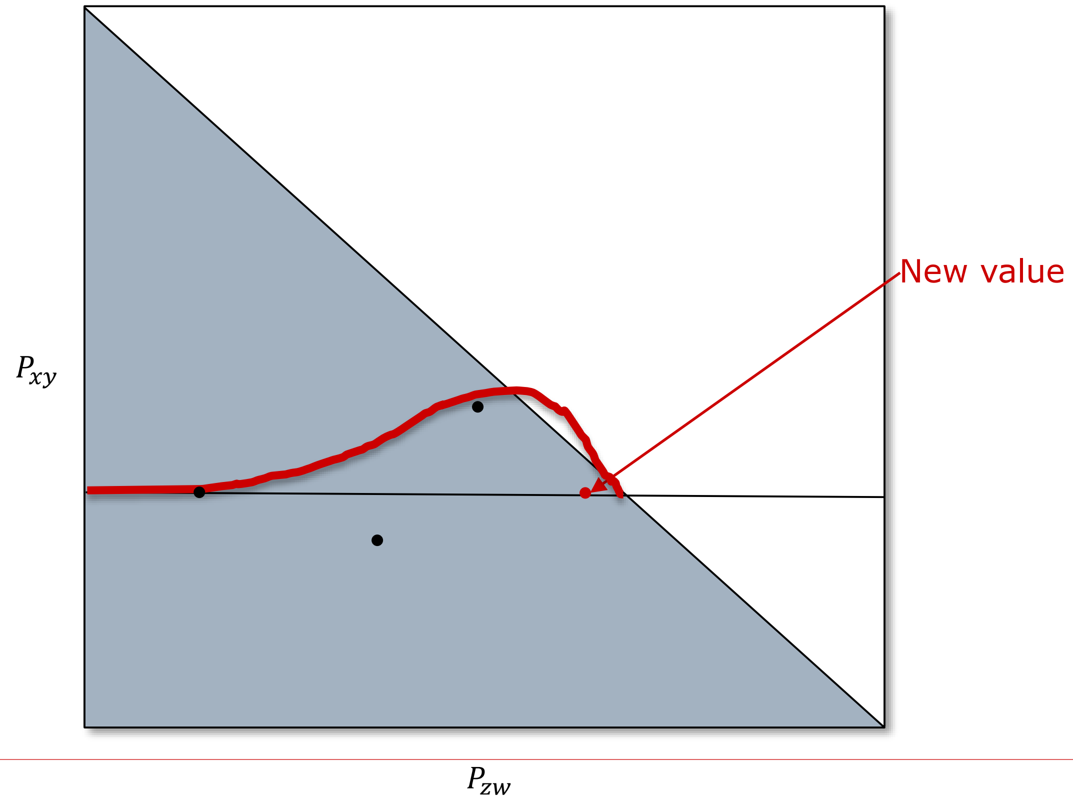


Illustration of constrained Gibbs sampler

- Sample a new value of P_{zw} conditioned on the new current value of P_{xy}.

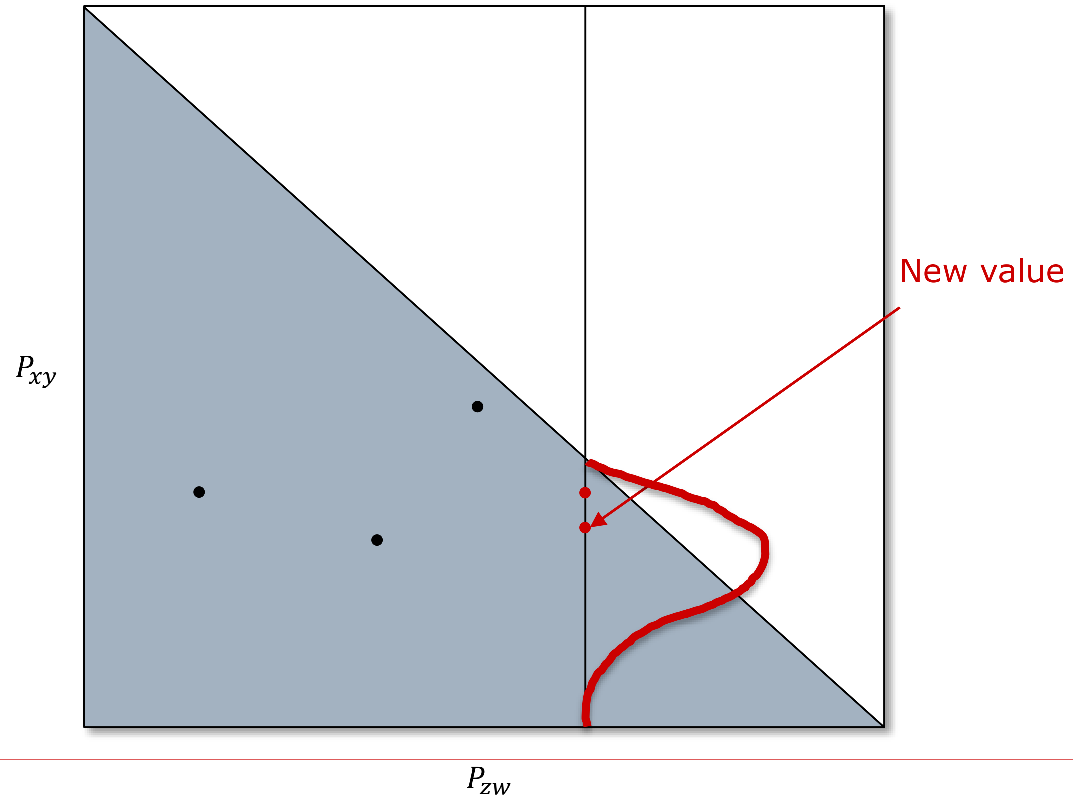


Illustration of constrained Gibbs sampler

- Keep the new pair: $\boldsymbol{P}^{(4)} = (P_{zw}^{(4)}, P_{xy}^{(4)})$.

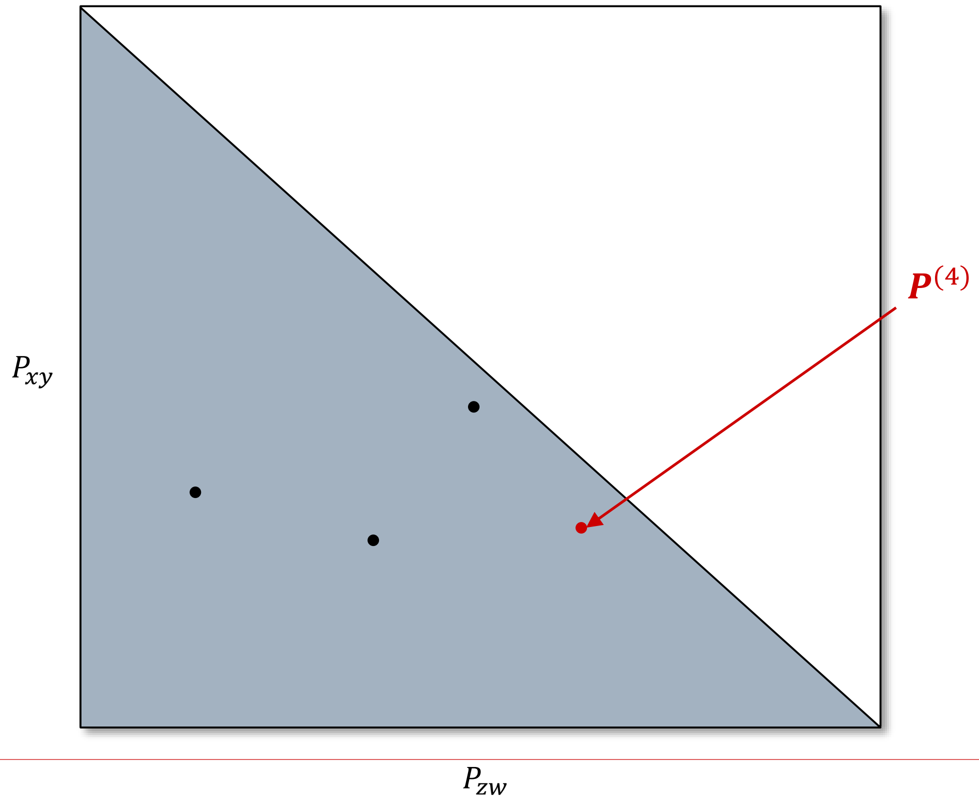
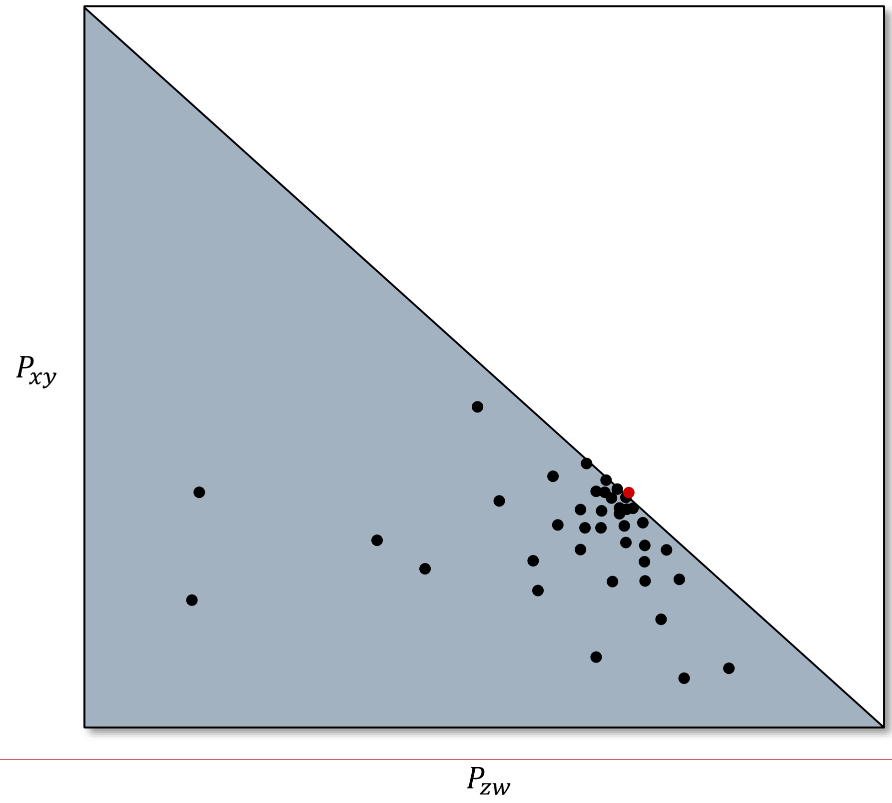


Illustration of constrained Gibbs sampler

- Repeat until convergence.



Summary

- Posterior moments can be calculated from the converged samples $\{\boldsymbol{P}^{(t)}; t = 1, \dots, T\}$
  - Posterior mean: $\overline{P_{xy}} = \frac{1}{T}\sum_{i=1}^{T} P_{xy}^{(t)}$
  - 95% Bayesian confidence interval

- There are <u>many</u> other methods for posterior simulation.
  - See Gelman et al. (2014) *Bayesian Data Analysis* for more methods and examples.

Model Evaluation

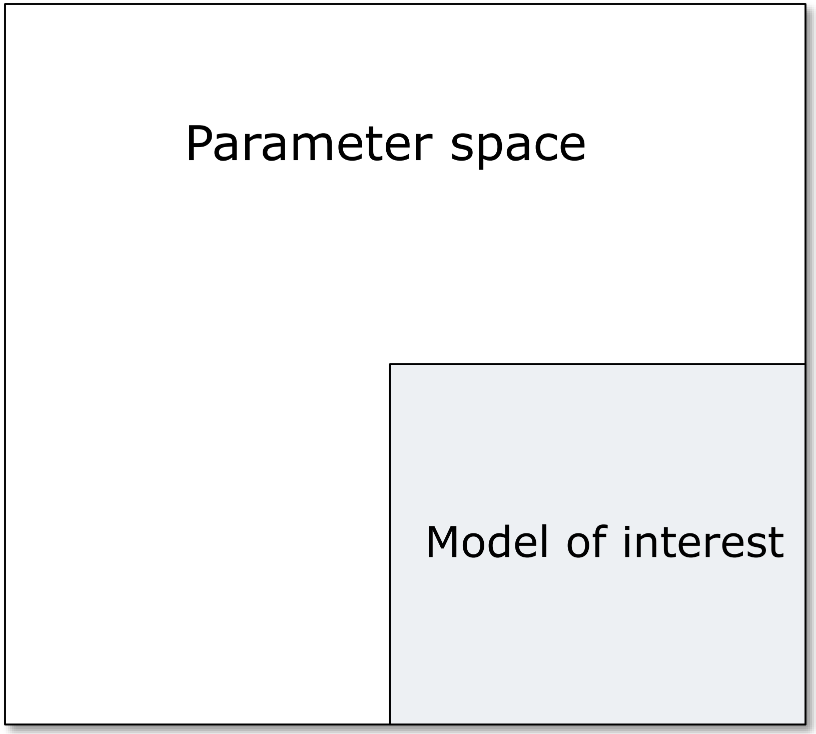
- Does the model provide an *adequate fit* to the data?

- *Triage* models to eliminate those that are inadequate.

- Bayesian p-value approach:
  - Does the discrepancy between the data and the model exceed what we would expect from binomial (sampling) error?

Model Evaluation

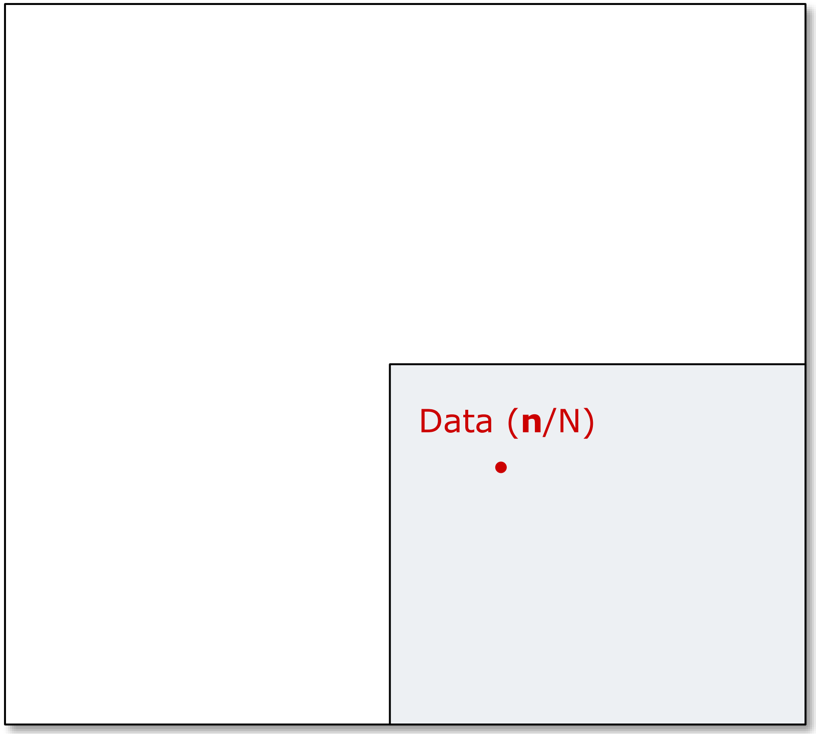
- Illustration



Model Evaluation

- Illustration

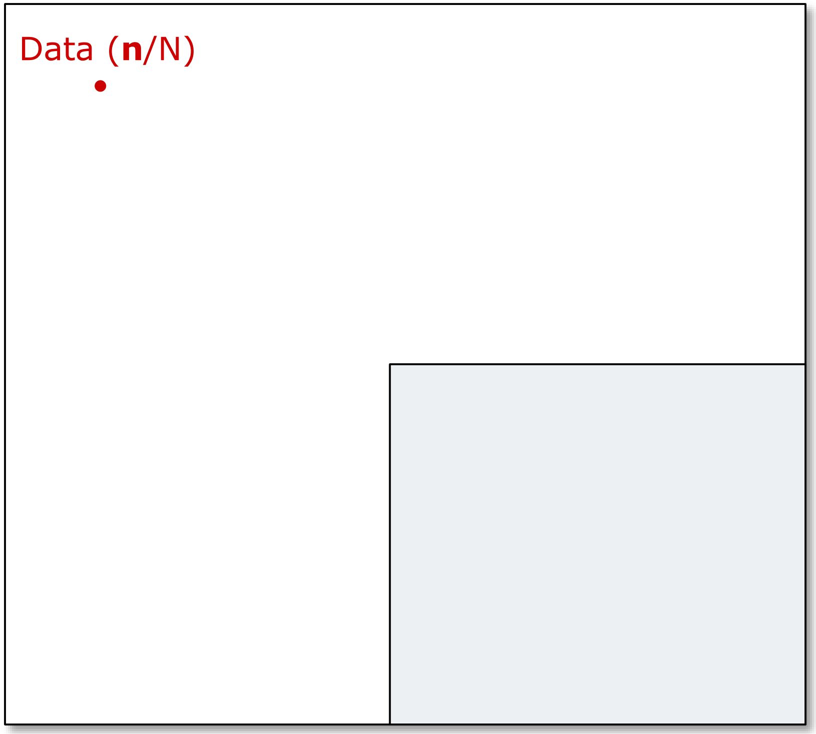
Perfect Fit!



Model Evaluation

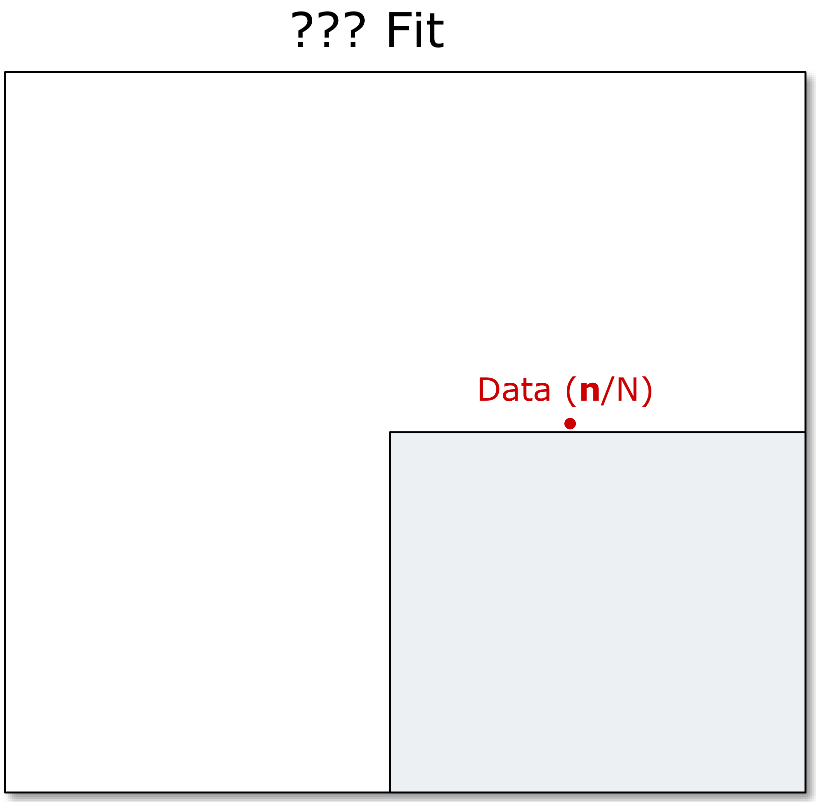
- Illustration

Bad Fit!



Model Evaluation

- Illustration



Does this discrepancy (between the data and the model)
exceed what we would expect from binomial (sampling) error?

Model Evaluation

- The *posterior predictive distribution* tells us how far the data should be from the model based on binomial error alone.

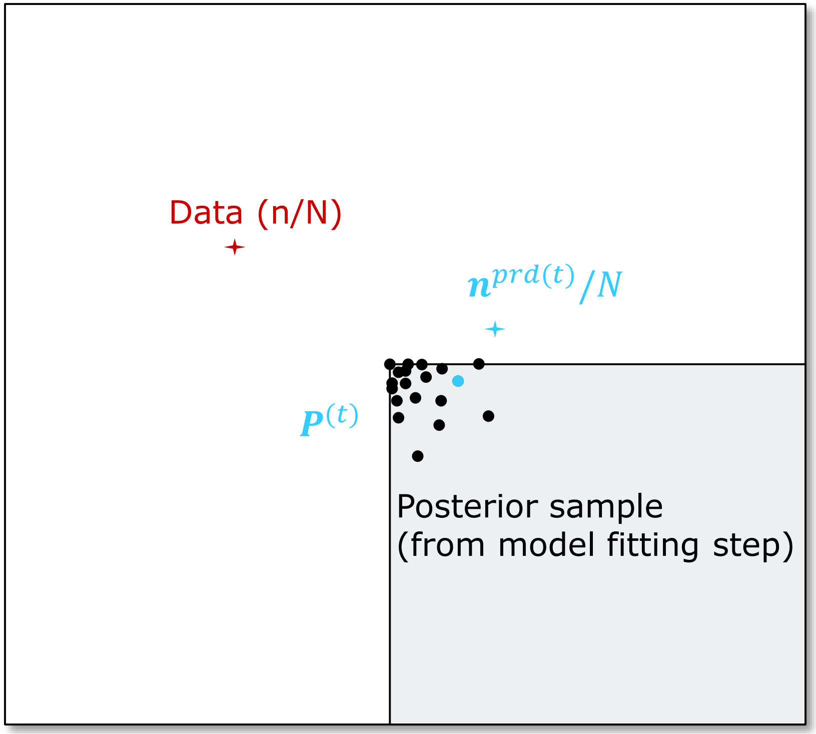
$$\pi\left(\boldsymbol{n}^{prd}\middle|\boldsymbol{n}, m\right) = \int_{\Lambda_m} p\left(\boldsymbol{n}^{prd}\middle|\boldsymbol{P}\right)\pi(\boldsymbol{P}|\boldsymbol{n})d\boldsymbol{P}$$

- Estimation:
  - For each parameter sample (from the model fitting step), draw a binomial random vector ($\boldsymbol{n}^{prd(t)}$) with probability of success given by that parameter draw.

- How do the posterior predictive samples compare to the actual data?

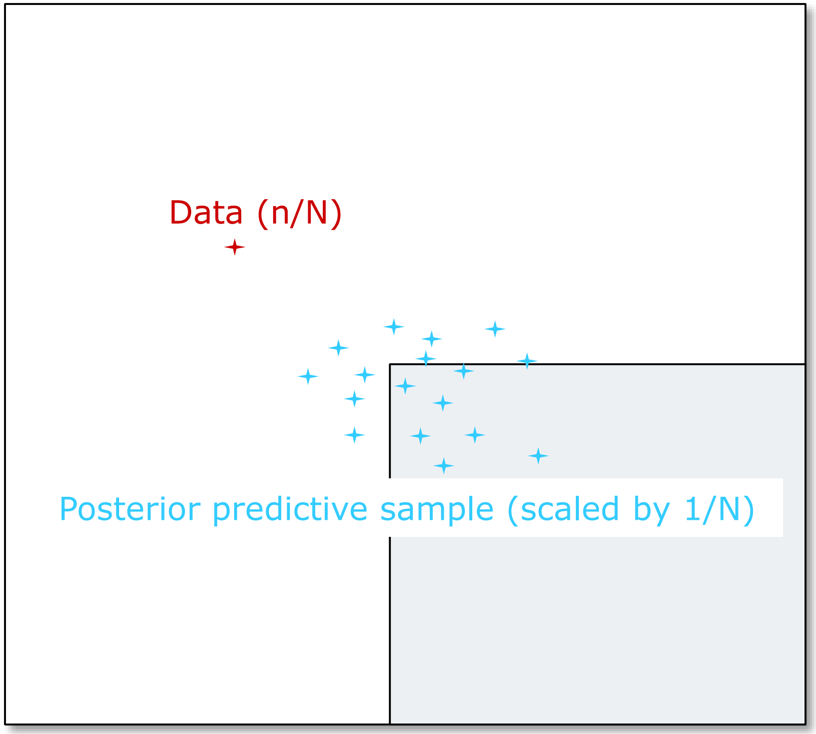
Bayesian p-value

- Illustration



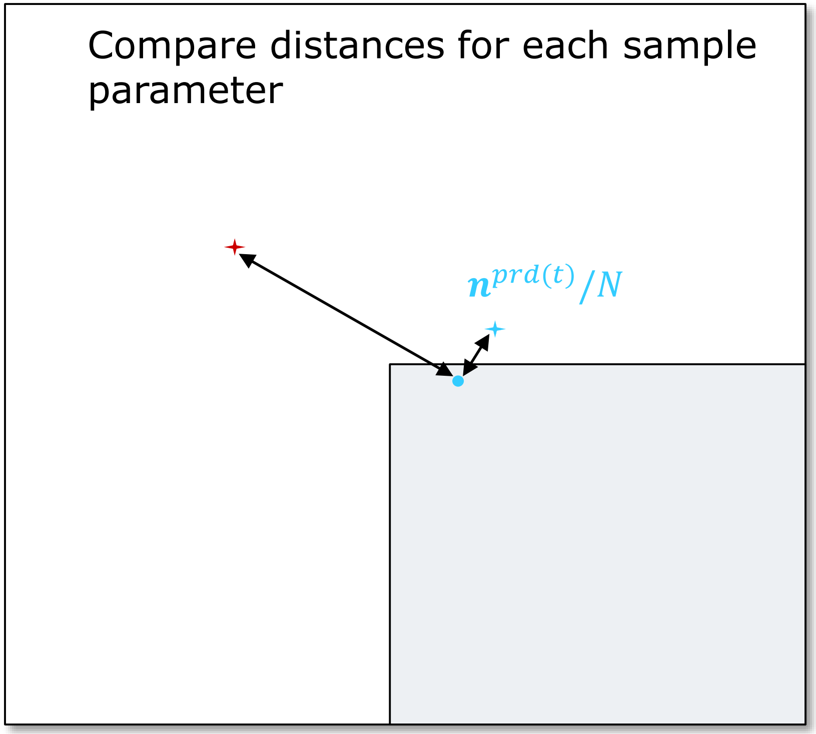
Bayesian p-value

- Illustration



Bayesian p-value

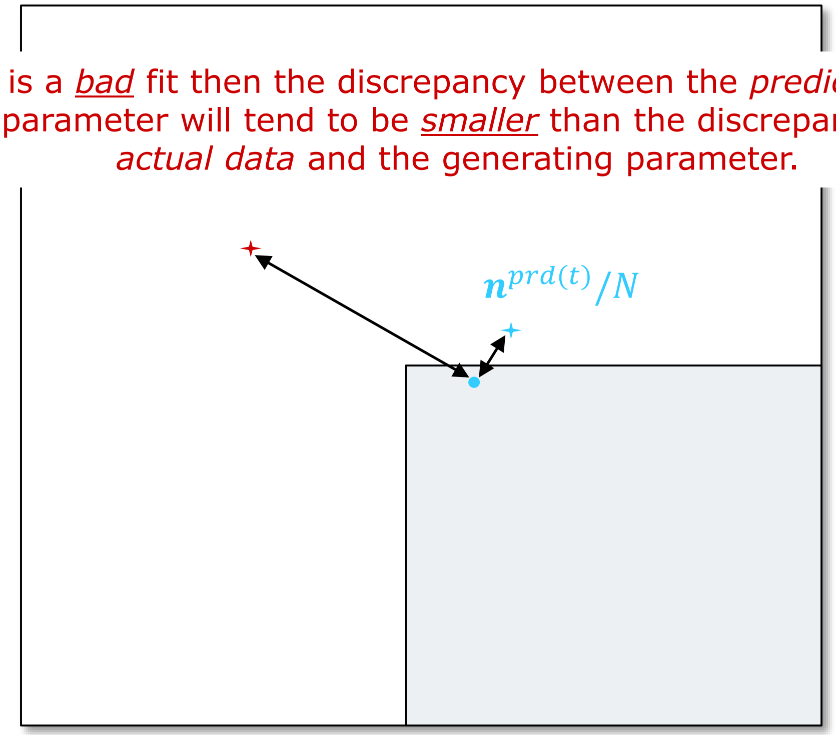
- Illustration



Bayesian p-value

- Illustration

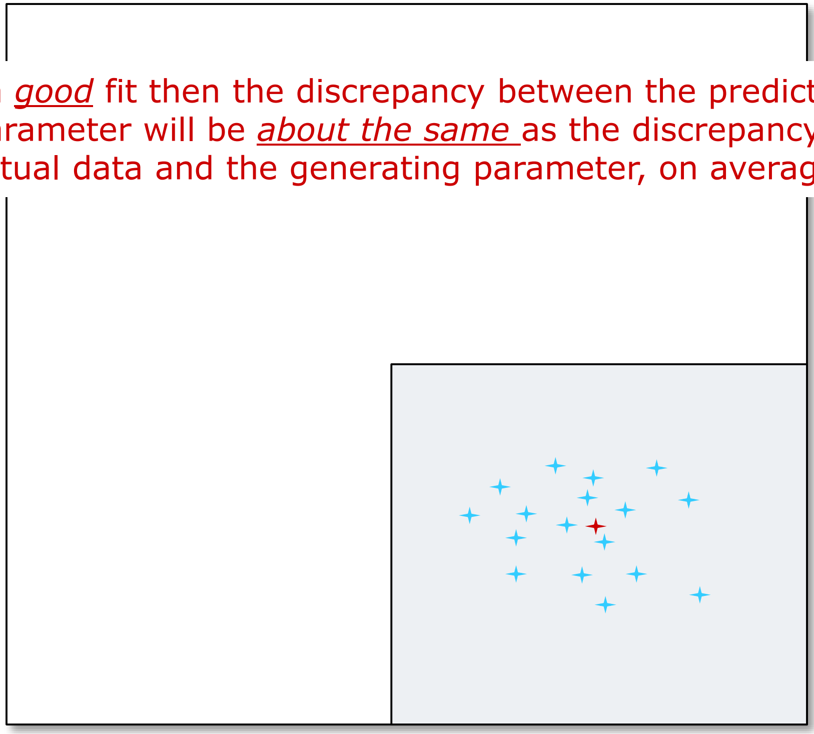
If the model is a <u>*bad*</u> fit then the discrepancy between the *predicted data* and the generating parameter will tend to be <u>*smaller*</u> than the discrepancy between the *actual data* and the generating parameter.



Bayesian p-value

- Illustration

If the model is a *good* fit then the discrepancy between the predicted data and the generating parameter will be *about the same* as the discrepancy between the actual data and the generating parameter, on average.



Bayesian p-value

- Compuation of Bayesian p-value:

$$p - \text{value} = \Pr\{\chi^2\left(\boldsymbol{n}^{prd}; \boldsymbol{P}\right) \geq \chi^2(\boldsymbol{n}; \boldsymbol{P})\}$$

  - A larger Bayesian p-value (e.g., near 0.5) indicates an adequate fit of the model to the data.
  - A low value (e.g., <.05) suggests a lack of fit.
    - In practice, this classification closely tracks the frequentist p-value.
  - A large Bayesian p-value does NOT imply that the form of the underlying data-generating process has been identified!

Model Selection

- Model selection involves a set of competing models, all of which have been found to provide a "good" fit to the data.
- Which model best approximates the underlying mental process?
- We must consider not only the goodness-of-fit of each model, but also their respective *complexity*.

Model Selection

- Complexity is typically *approximated* by counting parameters.
  - Akaike Information Criterion (AIC)

$$AIC = -2 \ln f\left(y \middle| \hat{\theta}\right) + 2k$$

  - Bayesian Information Criterion (BIC)

$$BIC = -2 \ln f\left(y \middle| \hat{\theta}\right) + k \ln n$$

- But what if the models under consideration have the same number of parameters!

Model Selection

- Deviance Information Criterion (DIC)
  - Posterior predictive approach to model selection.
  - Minimizes the expected loss in predicting a replicate of the observed data.
  - <u>Advantage</u>: Relatively easy to compute if you already have a sample from the from the posterior distribution (see Myung et al., 2005 for details).
  - <u>Disadvantage</u>: Only provides ordinal information (smaller is better) on the suitability of models under consideration.

Model Selection

- Bayes factor
  - Prior predictive approach to model selection.
    - Evaluates the model *before* it has been updated with data
  - Selects the model that is most likely, given the observed data.
  - Provides a readily interpretable metric of evidence based on the relative likelihood of each model.
  - Computationally intensive to compute in general, but can be implemented for order constrained models via the methodology of Klugkist and Hoijtink (2007).

Model Selection

- The Bayes factor for any model m_k relative to m_0 is defined as

$$BF_{k0} = \frac{p(\boldsymbol{n}|m_k)}{p(\boldsymbol{n}|m_0)} = \frac{\int p(\boldsymbol{n}|\boldsymbol{P}, m_k)\pi(\boldsymbol{P}|m_k)\, d\boldsymbol{P}}{\int p(\boldsymbol{n}|\boldsymbol{P}, m_0)\pi(\boldsymbol{P}|m_0)\, d\boldsymbol{P}}$$

- For nested models, the above formula simplifies to the ratio of two proportions

$$BF_{k0} = \frac{d_k}{c_k}$$

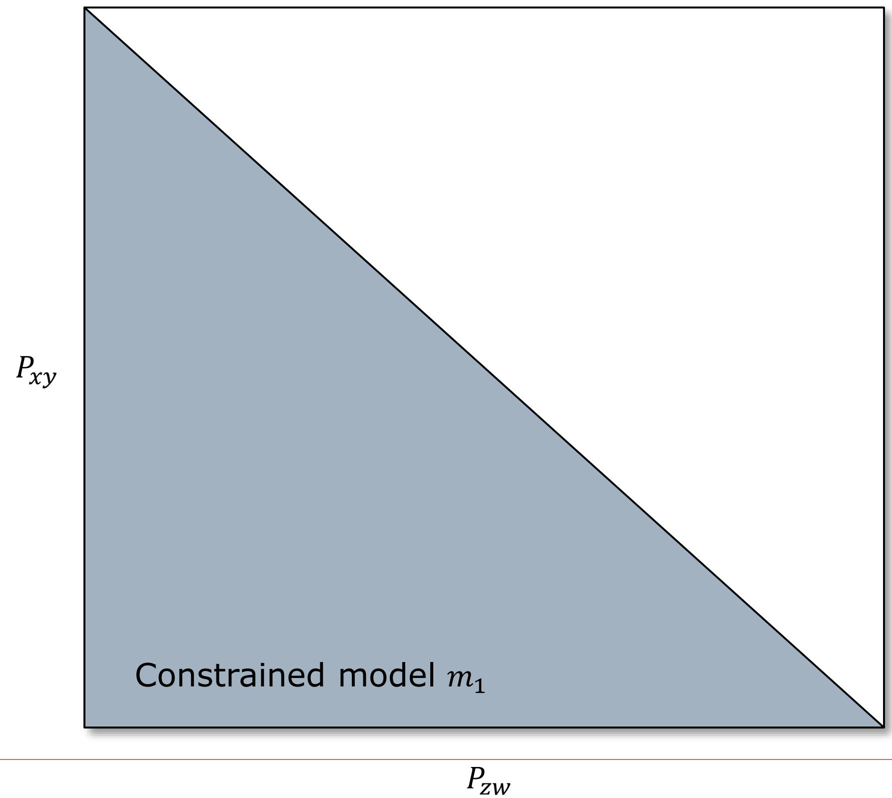
d_k: Proportion of the encompassing *posterior* in agreement with the constraints of m_k

c_k: Proportion of the encompassing *prior* in agreement with the constraints of m_k

$$BF_{23} = \frac{BF_{20}}{BF_{30}}$$

Model Selection

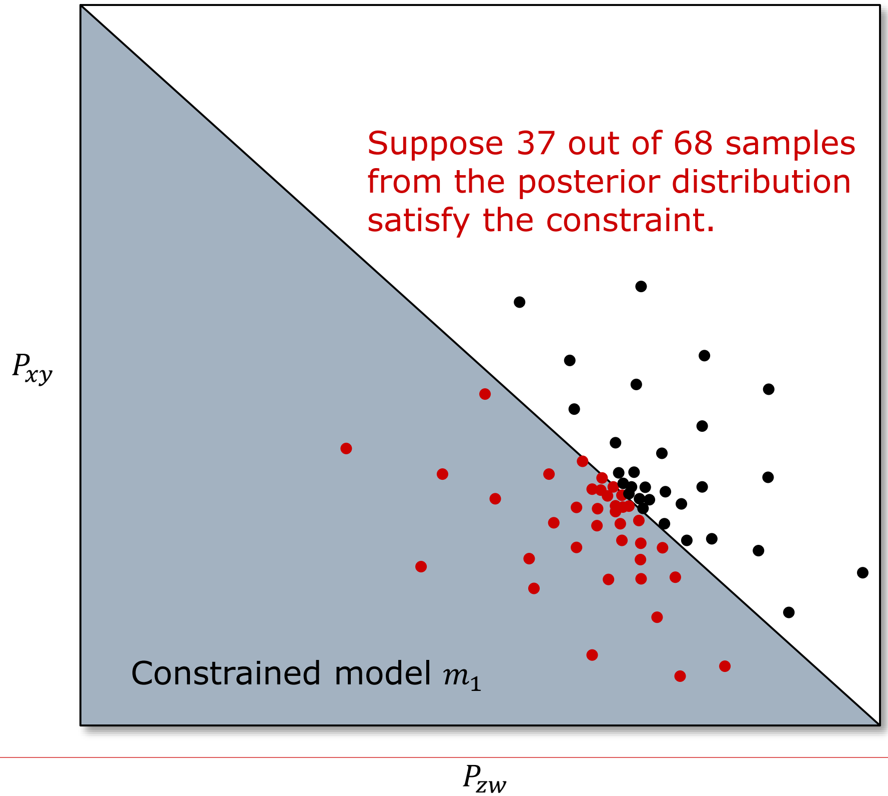
- Consider the model m_1 defined by $P_{xy} + P_{zw} < 1$



$$c_1 = \frac{1}{2}$$

Model Selection

- Consider the model m_1 defined by $P_{xy} + P_{zw} < 1$



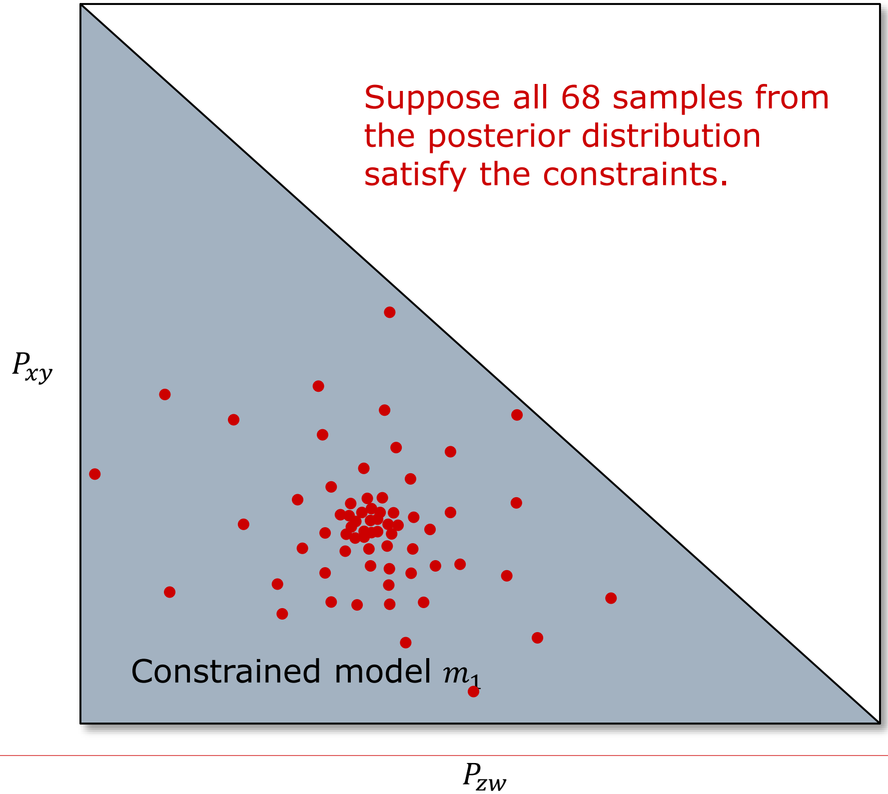
$$c_1 = \frac{1}{2}$$

$$d_1 = \frac{37}{68}$$

$$BF_{10} = \frac{37/68}{1/2} = 1.09$$

Model Selection

- Consider the model m_1 defined by $P_{xy} + P_{zw} < 1$



$$c_1 = \frac{1}{2}$$

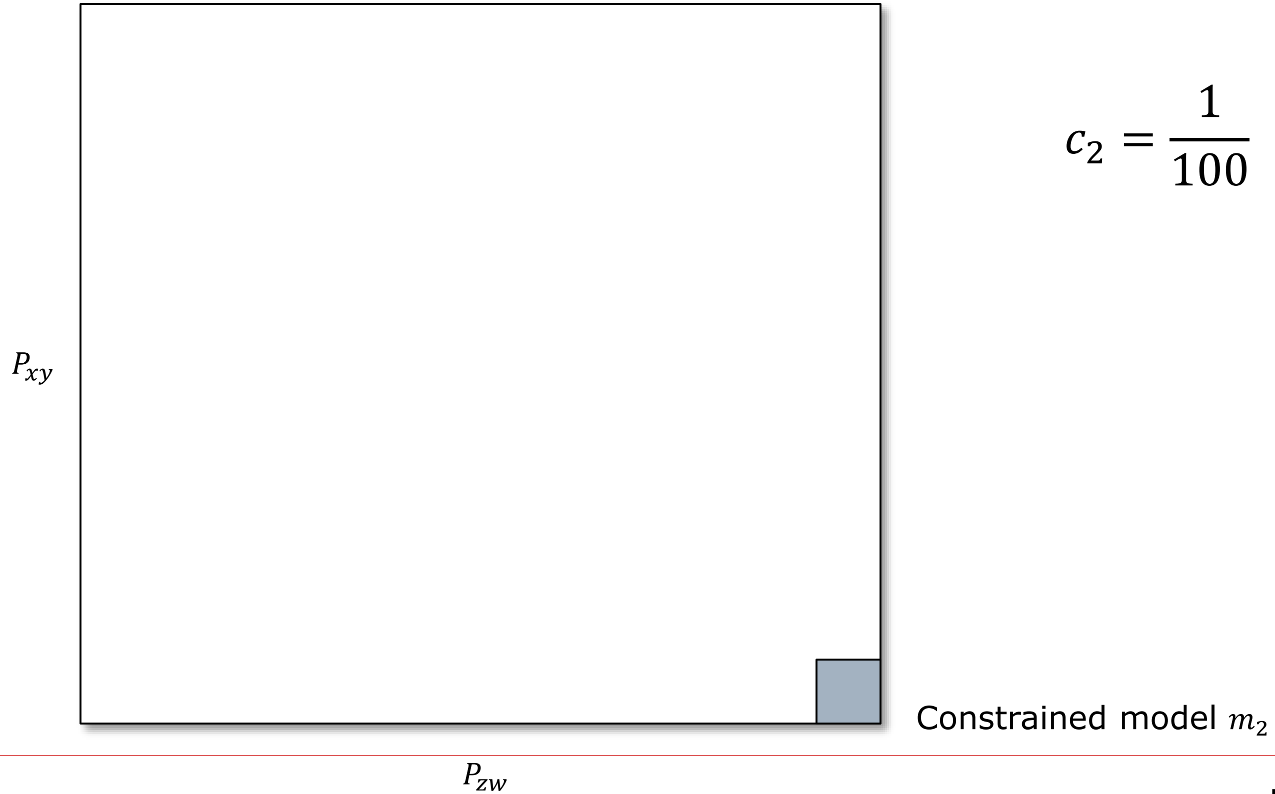
$$d_1 = \frac{68}{68}$$

$$BF_{10} = \frac{68/68}{1/2} = 2.00$$

The *largest* possible value of BF_{k0} is $\frac{1}{c_k}$

Model Selection

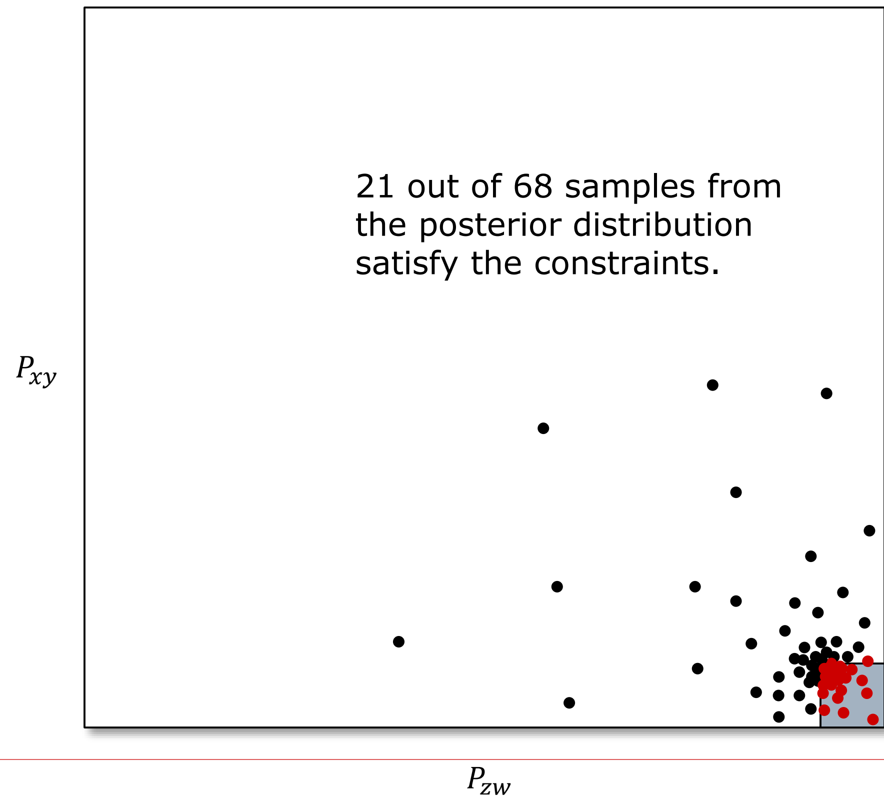
- Consider the model m_2 defined by $P_{xy} < 0.1;\ P_{ZW} > 0.9$



$$c_2 = \frac{1}{100}$$

Model Selection

- Consider the model m_2 defined by $P_{xy} < 0.1$; $P_{ZW} > 0.9$



$$c_2 = \frac{1}{100}$$

$$d_1 = \frac{21}{68}$$

$$BF_{10} = \frac{21/68}{1/100} = 30.88$$

The more constrained the model, the higher the Bayes factor!

Examples:

- Tversky (1969)
  - 8 subjects, 20 repeated choices on all pairwise combinations of 5 choice options (10 test pairs)

Subject 1

	M0	MMTP	WST	MST	SST
Bayes-p	--	0.32	0.029	0	0
Bayes Factor	**1.00**	0.11	0.00	0.00	0.00

Reanalysis by Cavagnaro and Davis-Stober (2014)

Examples:

- Tversky (1969)
  - 8 subjects, 20 repeated choices on all pairwise combinations of 5 choice options (10 test pairs)

Subject 7

	M0	MMTP	WST	MST	SST
Bayes-p	--	0.55	0.54	0.27	0.01
Bayes Factor	1.00	**16.24**	1.96	2.02	2.08

Reanalysis by Cavagnaro and Davis-Stober (2014)

Examples:

- Tversky (1969)
  - 8 subjects, 20 repeated choices on all pairwise combinations of 5 choice options (10 test pairs)

Subject 8

	M0	MMTP	WST	MST	SST
Bayes-p	--	0.54	0.54	0.56	0.12
Bayes Factor	1.00	17.96	6.67	41.65	**90.42**

Reanalysis by Cavagnaro and Davis-Stober (2014)

Outline

- ~~*Bayesian* Probabilistic Specification of a (Binary Choice) Decision Theory~~

- ~~*Bayesian* Quantitative testing of decision theories~~
  - ~~Model Estimation, Evaluation, and Selection~~

- Bayesian methods for evaluating group data

Group Data

- Pooled Bayes factor (PBF)
  - The PBF is the ratio of the marginal likelihoods of two models given the pooled data across all subjects.
  - The model with the highest PBF is the one that best accounts for the pooled data.
  - Computation is the same as for an individual, just pool the data.

Group Data

- Group Bayes factor (GBF)
  - The GBF is the product of the individual-level Bayes factors.
  - The model with the highest GBF is the one that *jointly* best accounts for every subject’s choice data.
  - A high GBF does *not* necessarily indicate that a model accounts well for every subject’s choice data, only that it does so better than other models.
  - May not allow sufficient heterogeneity

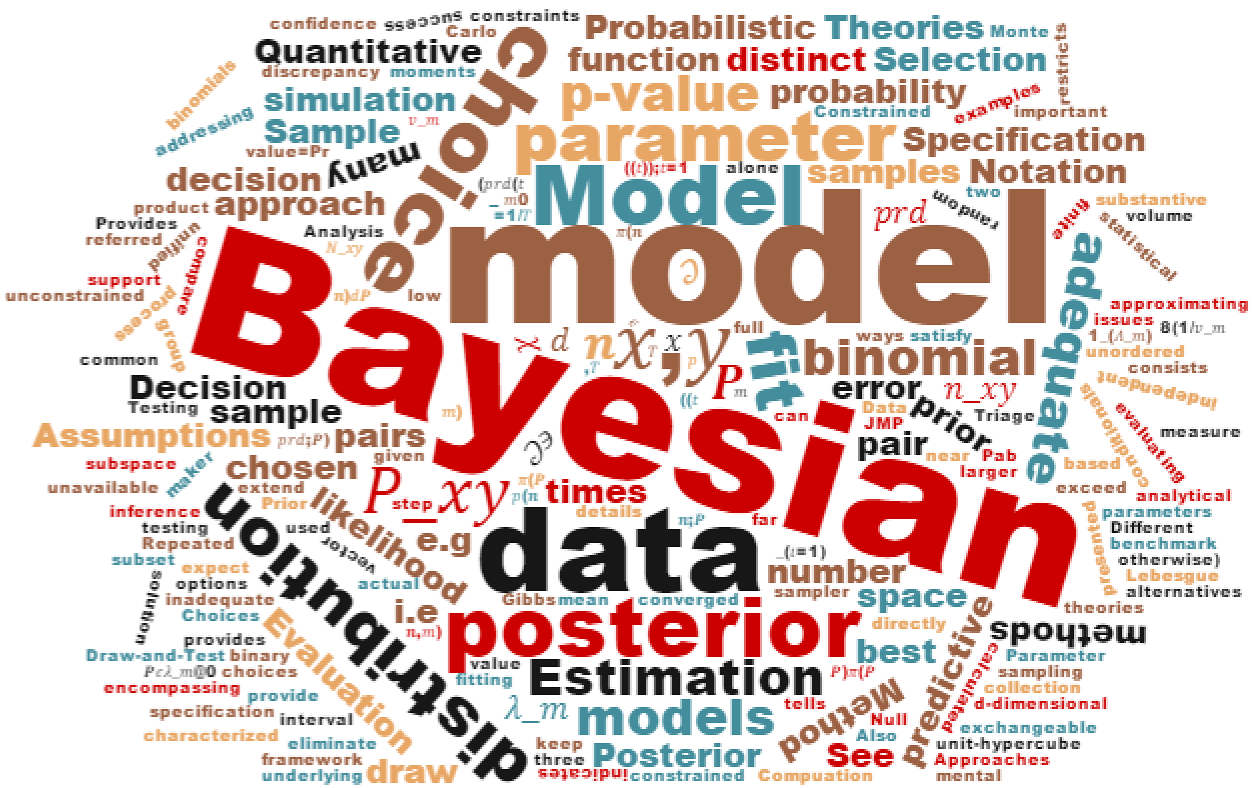
Group Data

- Hierarchical Bayesian Mixture Model
  - One can use the individual-level Bayes factors to estimate the proportion of subjects who are best described by each model.
  - Allows every subject to have their own model.
  - May result in overfitting!
  - Calculation boils down to arithmetic on the individual-level Bayes factors.
    - See Cavagnaro and Davis-Stober (Decision, 2014) for details calculation.

Summary

- The way to deal with group data depends on how much heterogeneity you wish to allow.
  - Assume every subject has the same binary choice probabilities:
    - Pooled Bayes factor (PBF).
  - Assume that binary choice probabilities satisfy the same set of constraints for every subject, but allow different subjects to have different choice probabilities within those constraints:
    - Group Bayes factor (GBF).
  - Allow every subject to have a set of constraints on their binary choice probabilities:
    - Hierarchical Bayesian mixture model.

The End



dcavagnaro@fullerton.edu
http://mihaylofaculty.fullerton.edu/sites/dcavagnaro