NOAA FISHERIES

Guidelines to validating generalized linear mixed models in Template Model Builder using quantile residuals

Andrea Havron[1], Cole Monnahan[2]
Florian Hartig[3], Kasper Kristensen[4], James Thorson[2]

[1]ECS Federal in support of NOAA Fisheries
[2]Alaska Fisheries Science Center
[3]University of Regensburg
[4]DTU Comput

Relevancy to Fishery Stock Assessments

- Stock Assessment models that use TMB:
  - WHAM
  - SAM
  - LIME
- Next Gen Stock Assessment Project: FIMS
  - TMB back-end
  - Random effects functionalities
- Need to establish clear guidelines

Why Validate?

Model Selection

- Which model is best?
- Tools: AIC, BIC
- Issues: degrees of freedom

Model Validation

- Do the data meet model assumptions?
- Tools: Residuals
- Issues: Variance

Pearson residuals

$$r_i = \frac{Y_i - E[Y_i]}{\sqrt{Var[Y_i]}}$$

GLMM Issues:

- Assumes the mean-variance relationship is equal to 1
- Correctly specified model invalidated more than expected
- Difficult to assess heteroscadacity and overdispersion

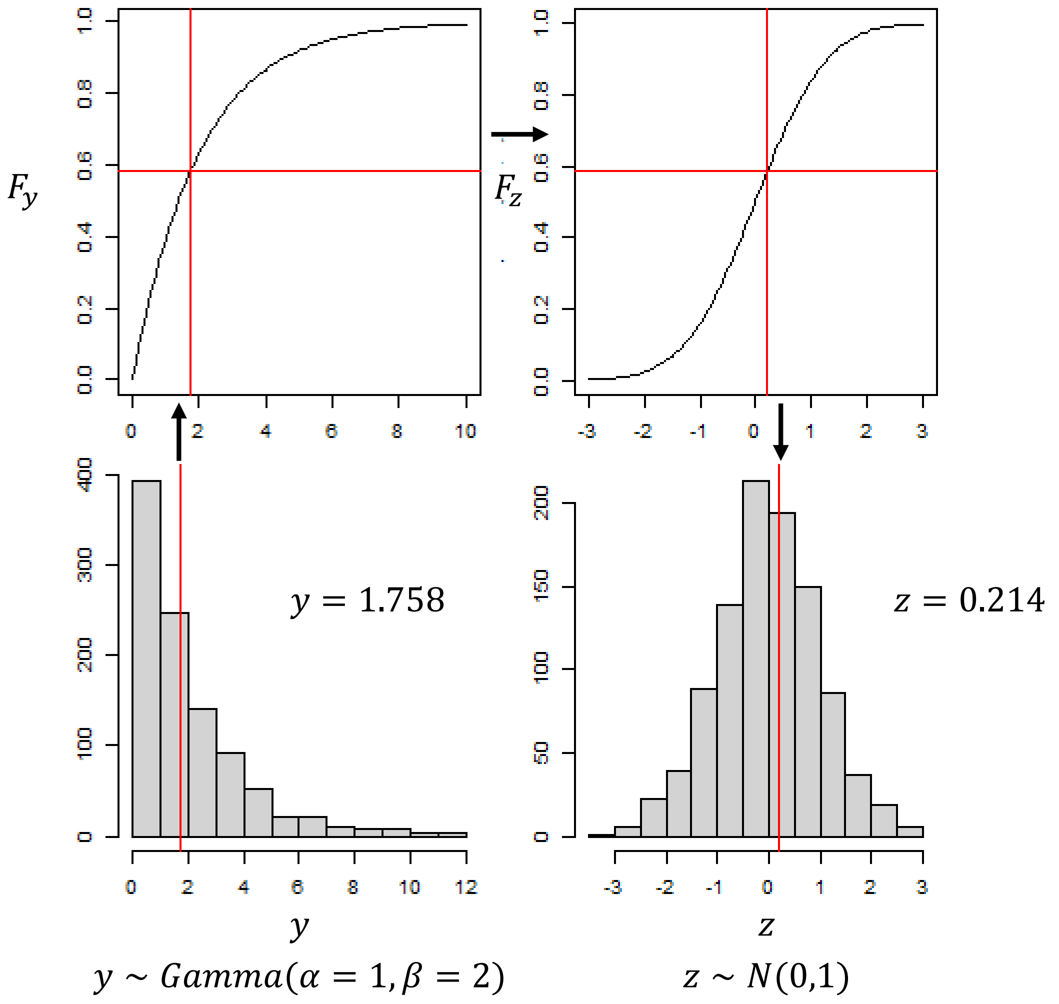
Quantile Residuals

$$r_i = \varphi^{-1}\{F(Y_i, \Theta)\}$$

φ^{-1}: cdf inverse of the standard normal distribution
$F(Y, \Theta)$: cdf of f(y, Θ)

Dunn & Smyth, 1996.

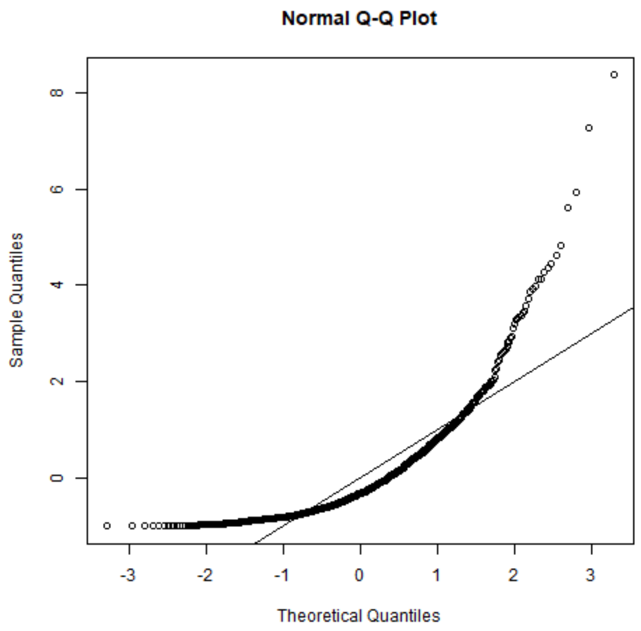
$F_y(y = 1.758) = 0.58 = F_z(z = 0.214)$



GOF: Standardizing to N(0,1) or U(0,1)

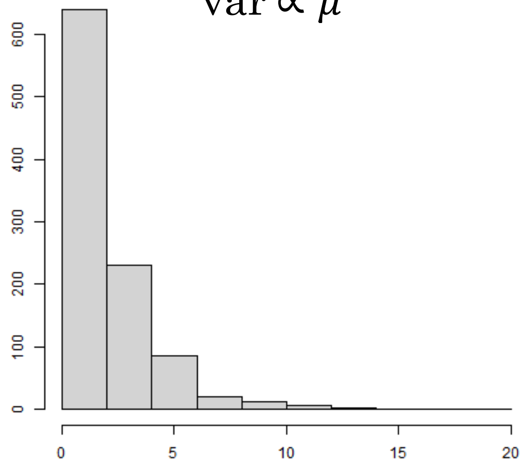
$$r_i = \frac{Y_i - E[Y_i]}{\sqrt{Var[Y_i]}}$$

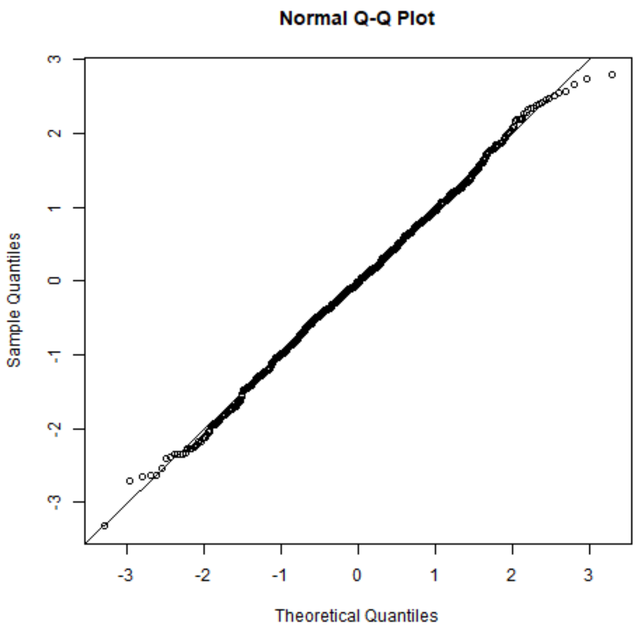
$$r_i = \varphi^{-1}\{F(Y_i, \Theta)\}$$



$y \sim Gamma(\alpha = 1, \beta = 2)$

$\mathrm{var} \propto \mu^2$





Quantile Residuals and GLMMs

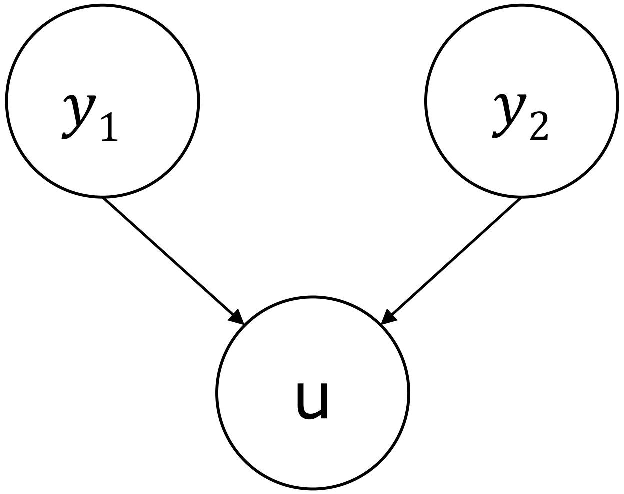
Considerations

- Need closed form solution to cdf
- Need to rotate multivariate to independent univariate (eg. temporal, spatial)

$$f(y;\theta) = \int_{\mathbb{R}} f(y|u;\theta) f(u;\theta) du$$

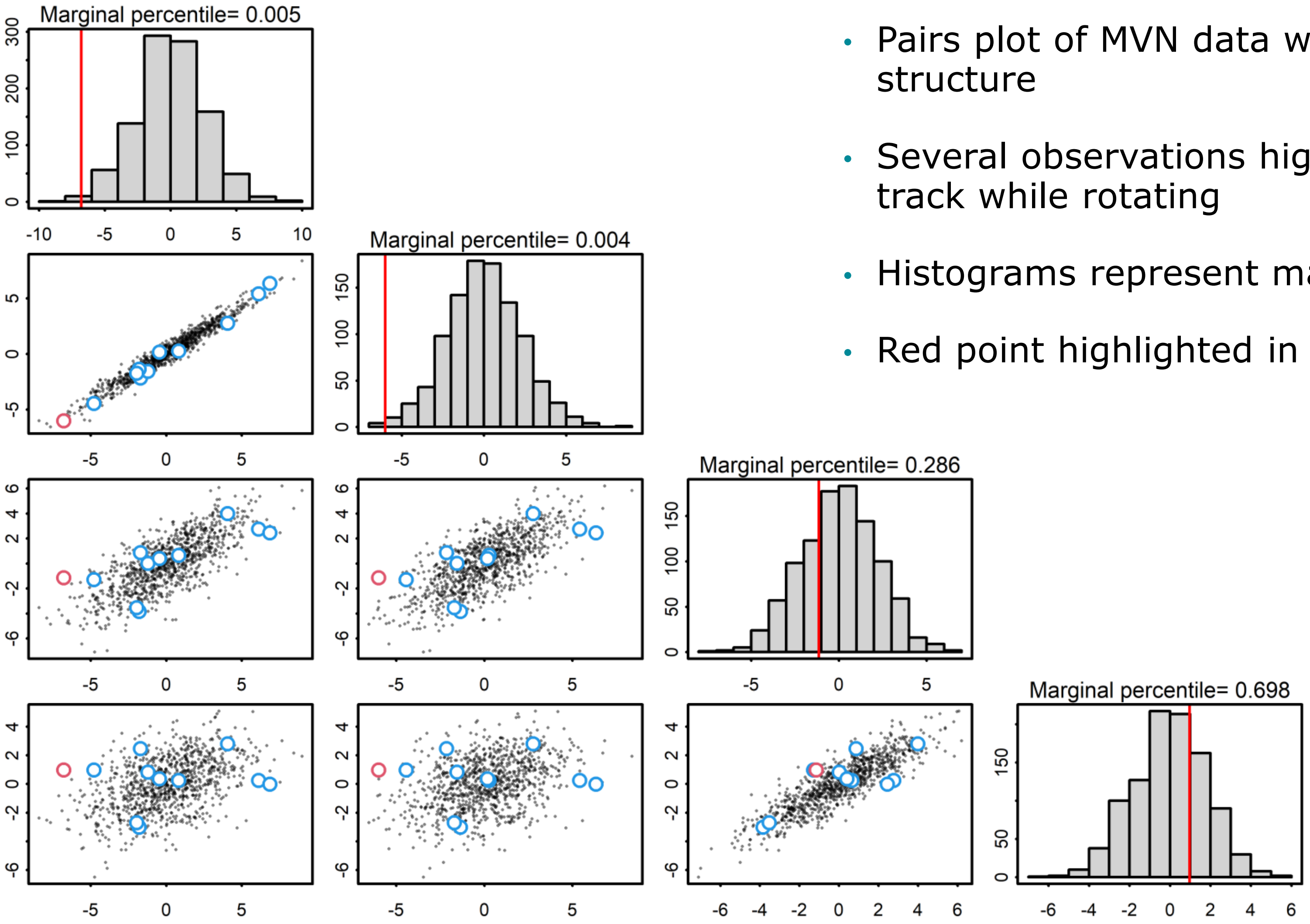
Conditional Independence

$$y_1 \perp y_2 \mid u$$



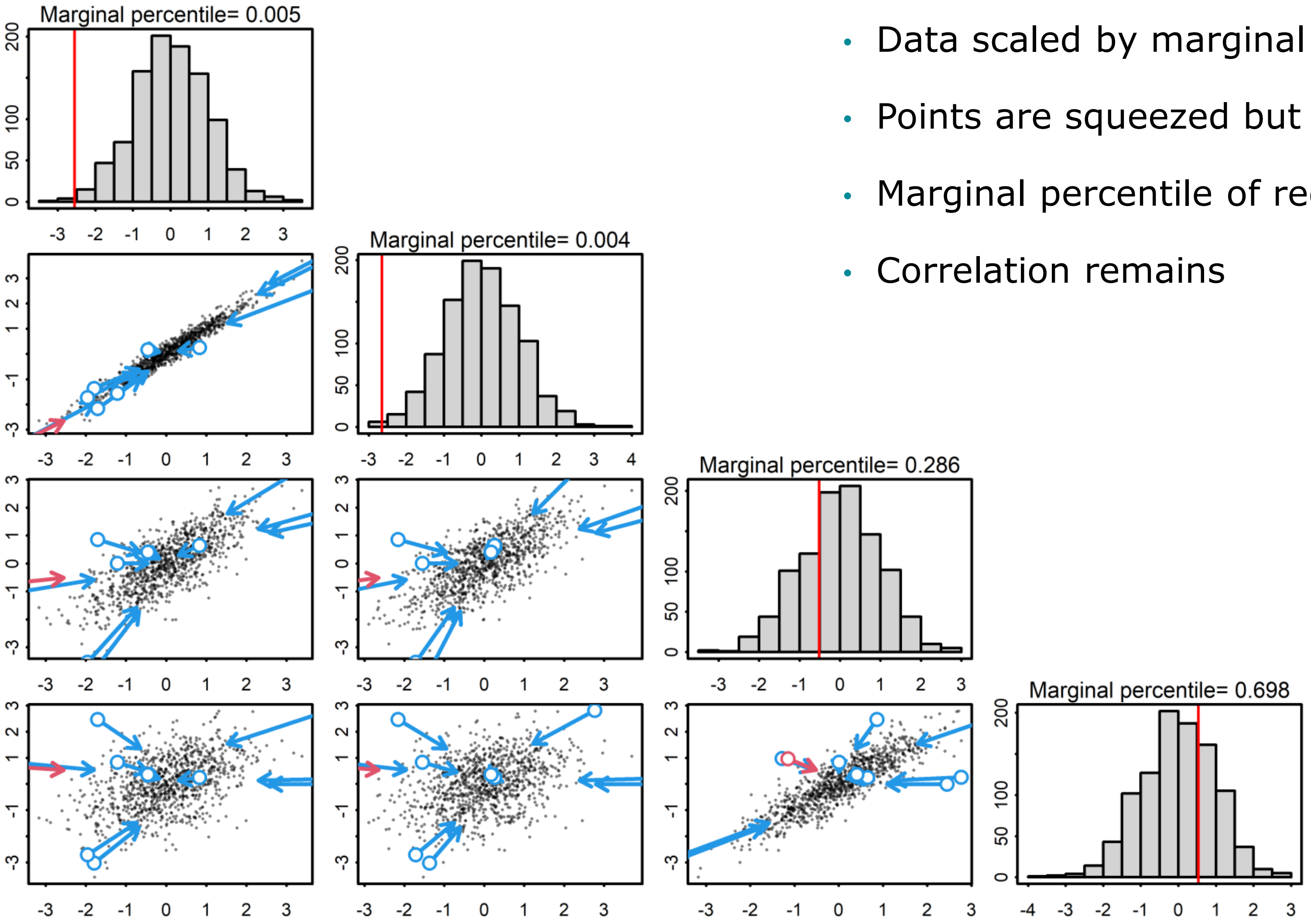
$$f(\boldsymbol{y}|u;\theta)f(u;\theta)$$

Correlated MVN



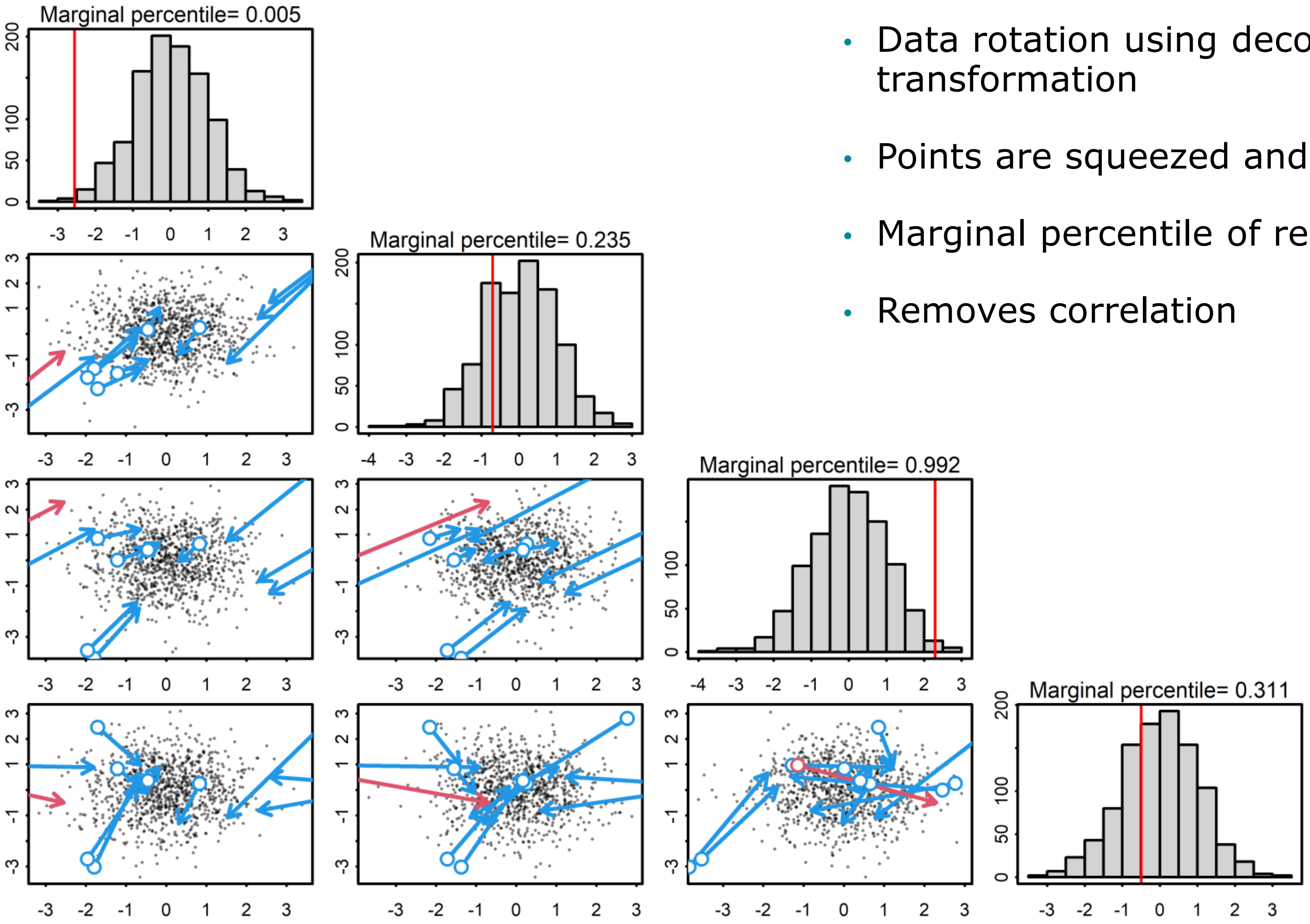
- Pairs plot of MVN data with correlation structure
- Several observations highlighted in order to track while rotating
- Histograms represent marginal distributions
- Red point highlighted in the histogram

Scaled to unit variance



- Data scaled by marginal variance
- Points are squeezed but not rotated
- Marginal percentile of red point unaffected
- Correlation remains

Scaled and Rotated



- Data rotation using decorrelation transformation
- Points are squeezed and rotated
- Marginal percentile of red point changes
- Removes correlation

Approximations to the Quantile Residual

TMB

- Analytical calculations
- Laplace approximations of the cdf
- Bayesian approximations of the cdf

Thygesen, U. et al., 2017.

DHARMa

- Simulation based approximations

Hartig, F. 2020.

TMB methods

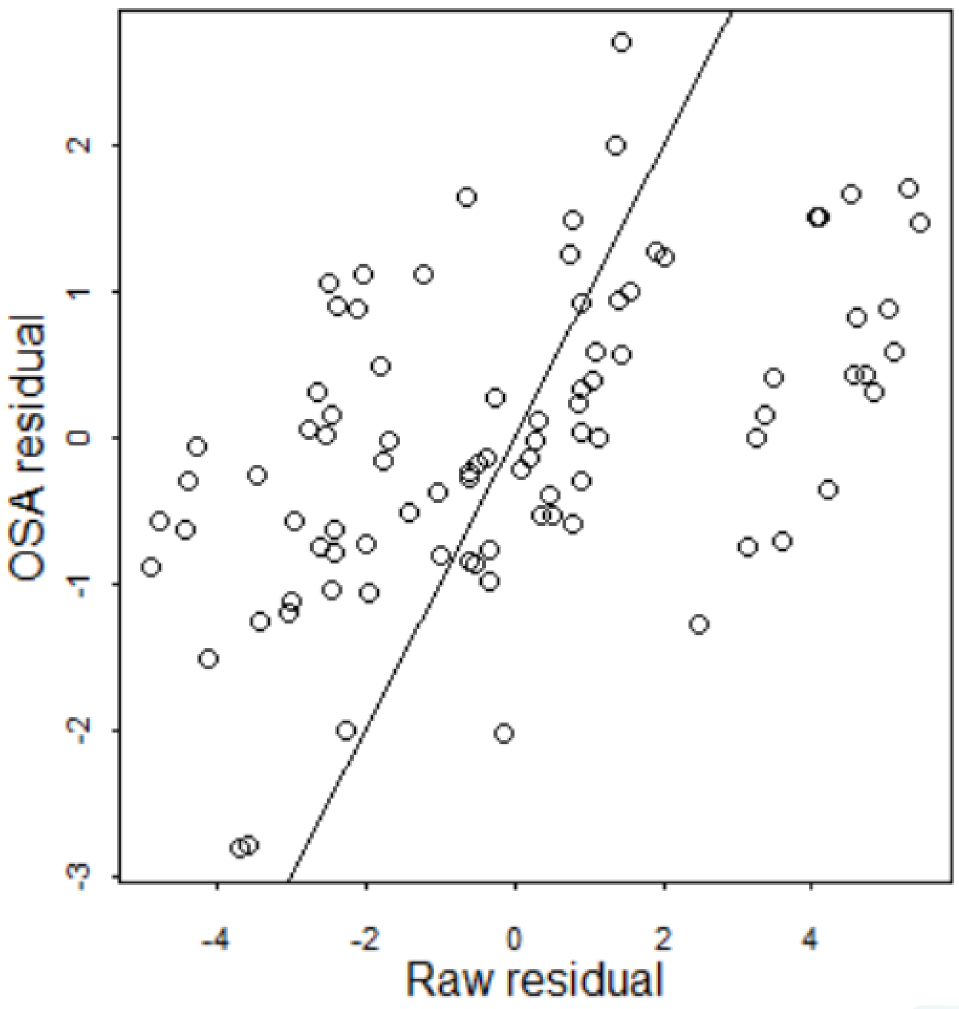
Full Gaussian (FG)	Assumes joint distribution of data and random effects is Gaussian. Applies rotation to transform to univariate space.
OneStepGaussian (OSG)	One-step conditional distribution is approximated with a Gaussian
CDF	One-step conditional distribution is a ratio of Laplace approximations, applied to the log(cdf)
MCMC	Fixed parameters are mapped to MLEs and tmbstan is used to draw a posterior of the random effects

TMB: Full Gaussian Method

- Analytical calculation of quantile residuals when observations are normal

<u>Simplified algorithm:</u>

1. Build `objnew` by adding data to 'random'
2. Call `objnew$fn()` to do inner optimization to get mode of expected data $\hat{y}$ (and RE) given FE
3. Apply fancy linear algebra to Hessian to get:
   $y \sim MVN(\hat{y}, \Sigma)$.
4. Rotate raw residuals to get OSA residuals:
   $r = chol(\Sigma) * (y - \hat{y})$



TMB One-Step Method

- One-step conditional approach using Laplace approximation

<u>Simplified algorithm:</u>

1. Map fixed parameters and random effects to MLEs
2. Iteratively add observation to subset y_1
3. Treat the rest of data as random effects and estimate using the Laplace approximation.
4. Calculate the residual as a ratio of the cdf of the subset to the cdf of the subset plus the cdf of the Laplace approximated data

$$U_i = \frac{P^M(Y_i \le y_i, Y_1^{i-1} = y_1^{i-1})}{P^M(Y_1^{i-1} = y_1^{i-1})}$$

$$Y_1^i = (Y_1, \dots, Y_i)$$

TMB MCMC Method

- Bayesian simulation approach

Simplified algorithm:

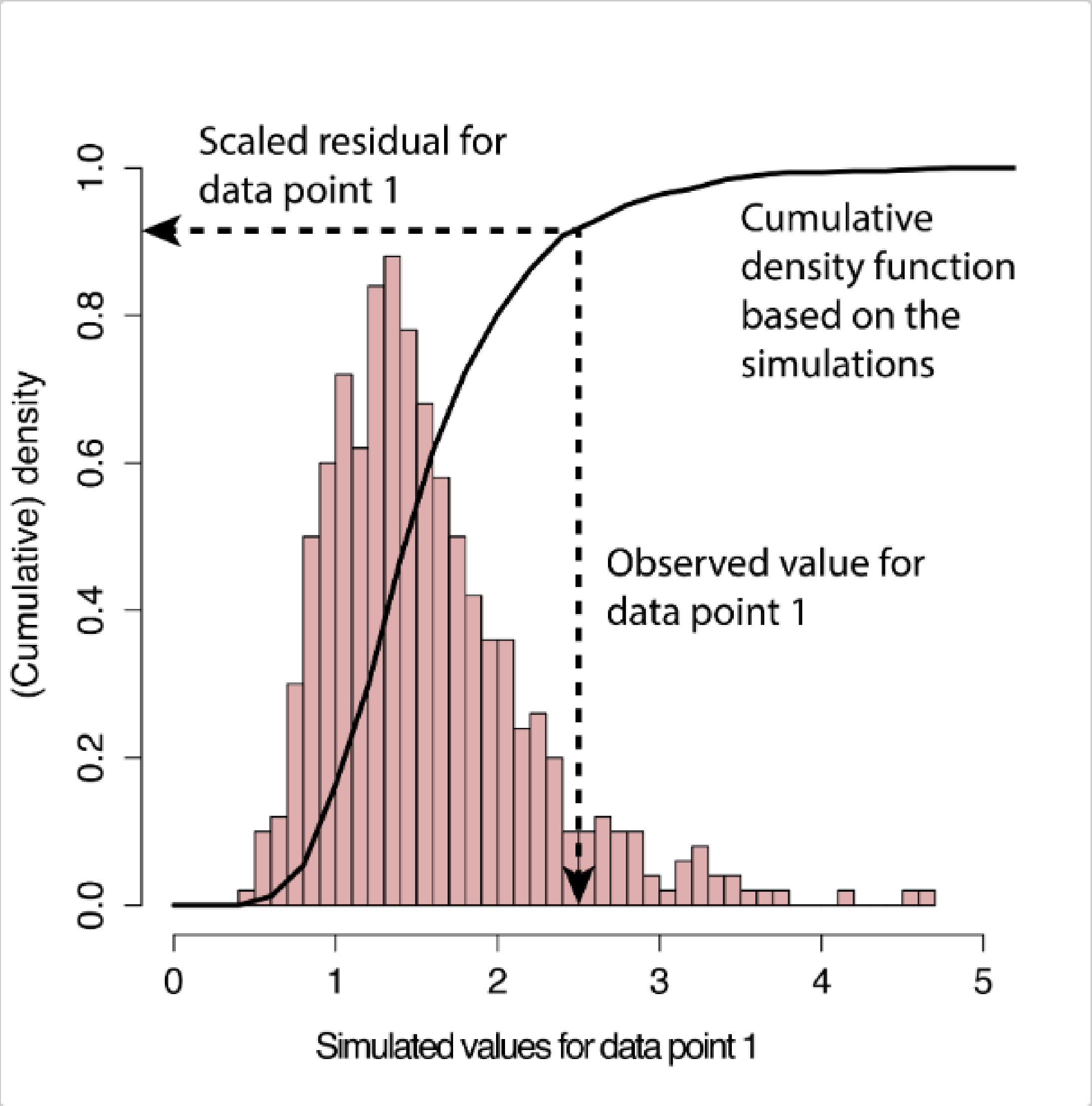
1. Map fixed parameters to their MLEs
2. Create a new object
3. Draw a single posterior from an MCMC chain
4. Use the posterior random vector to recalculate the expected value and plug into cdf calculations
5. Relies on conditional independence rather than rotation

```
obj2 <- MakeADFun(data, MLEs, map)
fit <- tmbstan(obj2,iter=warmup+1)
sample <- extract( fit )$u
mu <- beta0 + u
r <- qnorm( y, mu, sd )
```

DHARMa methods

Conditional	Simulate new observations conditional on the fitted random effects
Unconditional	Simulate new observations given new simulated random effects

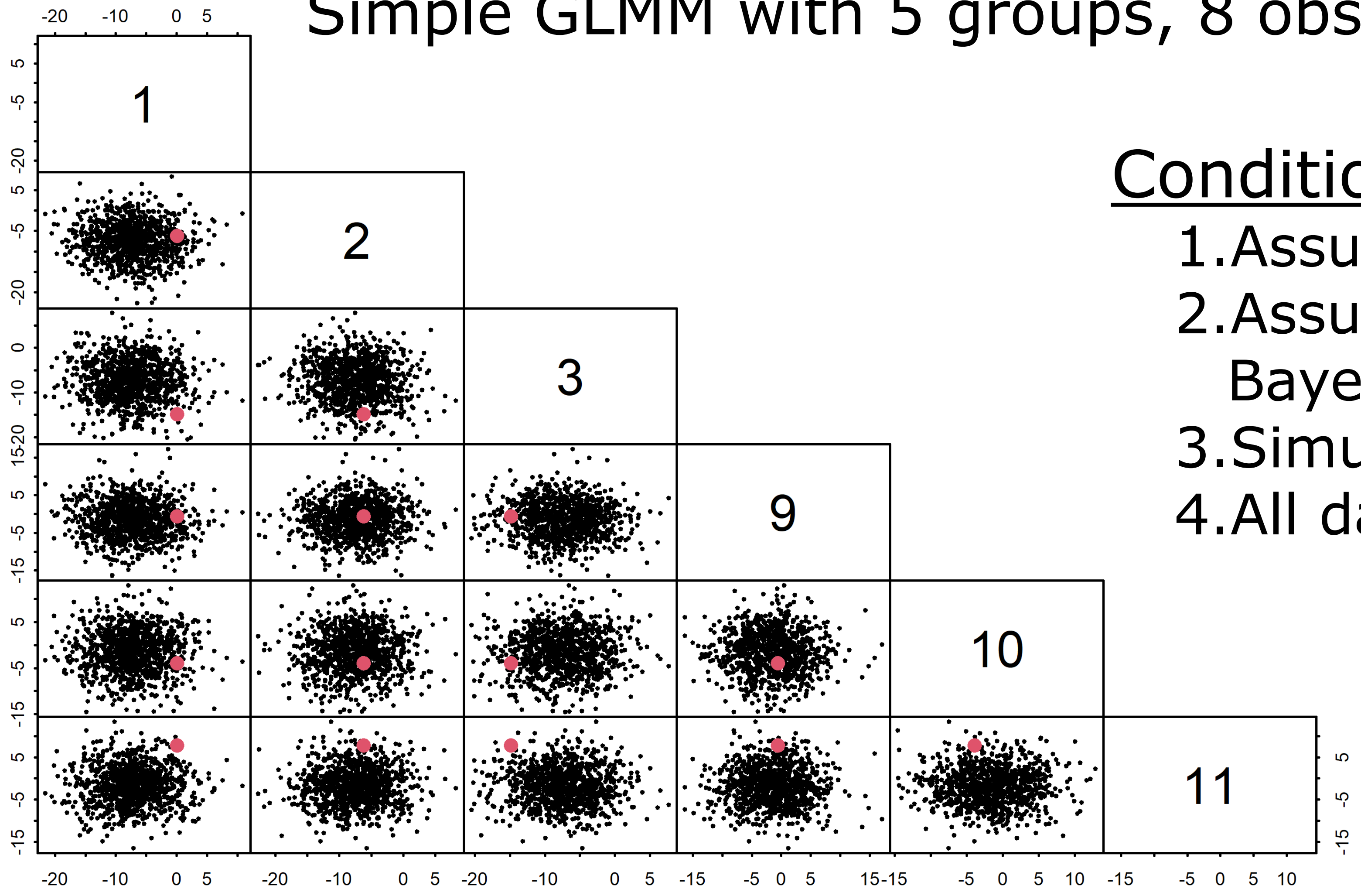
DHARMa: the empirical cdf



Source: Florence Hartig, DHARMa package

Conditional simulation example

Simple GLMM with 5 groups, 8 observations each

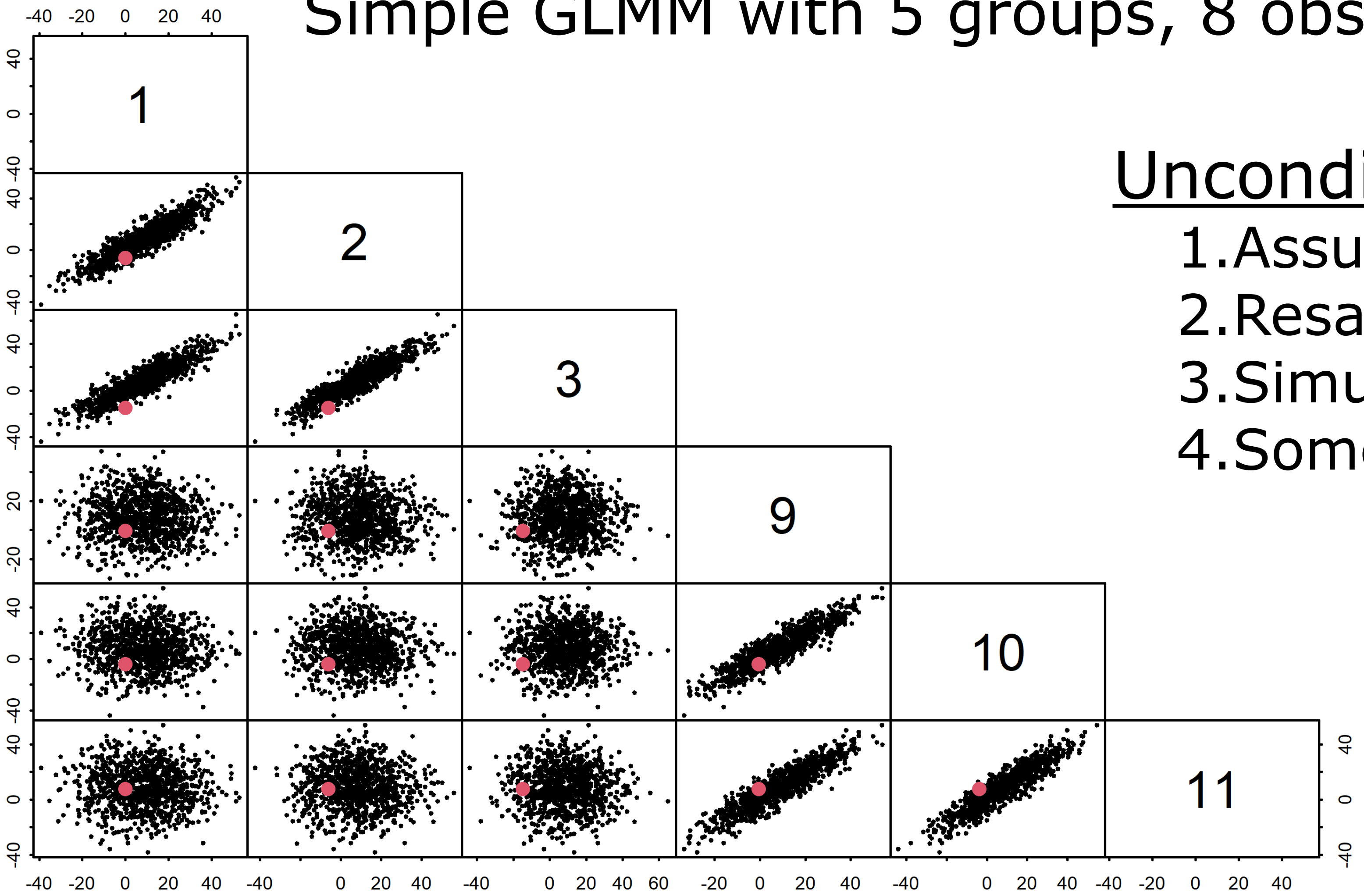


Conditional simulation:

1. Assume FE are the MLEs
2. Assume RE are the empirical Bayes estimates
3. Simulate a new data set
4. All data points uncorrelated

Unconditional simulation example

Simple GLMM with 5 groups, 8 observations each



Unconditional simulation:

1. Assume FE are the MLEs
2. Resample RE given FE
3. Simulate a new data set
4. Some data points correlated

Simulation Study

- Establish baseline behavior for different quantile residual calculation methods
- Perform GOF tests on correctly specified and mis-specified models
- Develop guidelines around when each method is or is not appropriate to use

Simulation Models

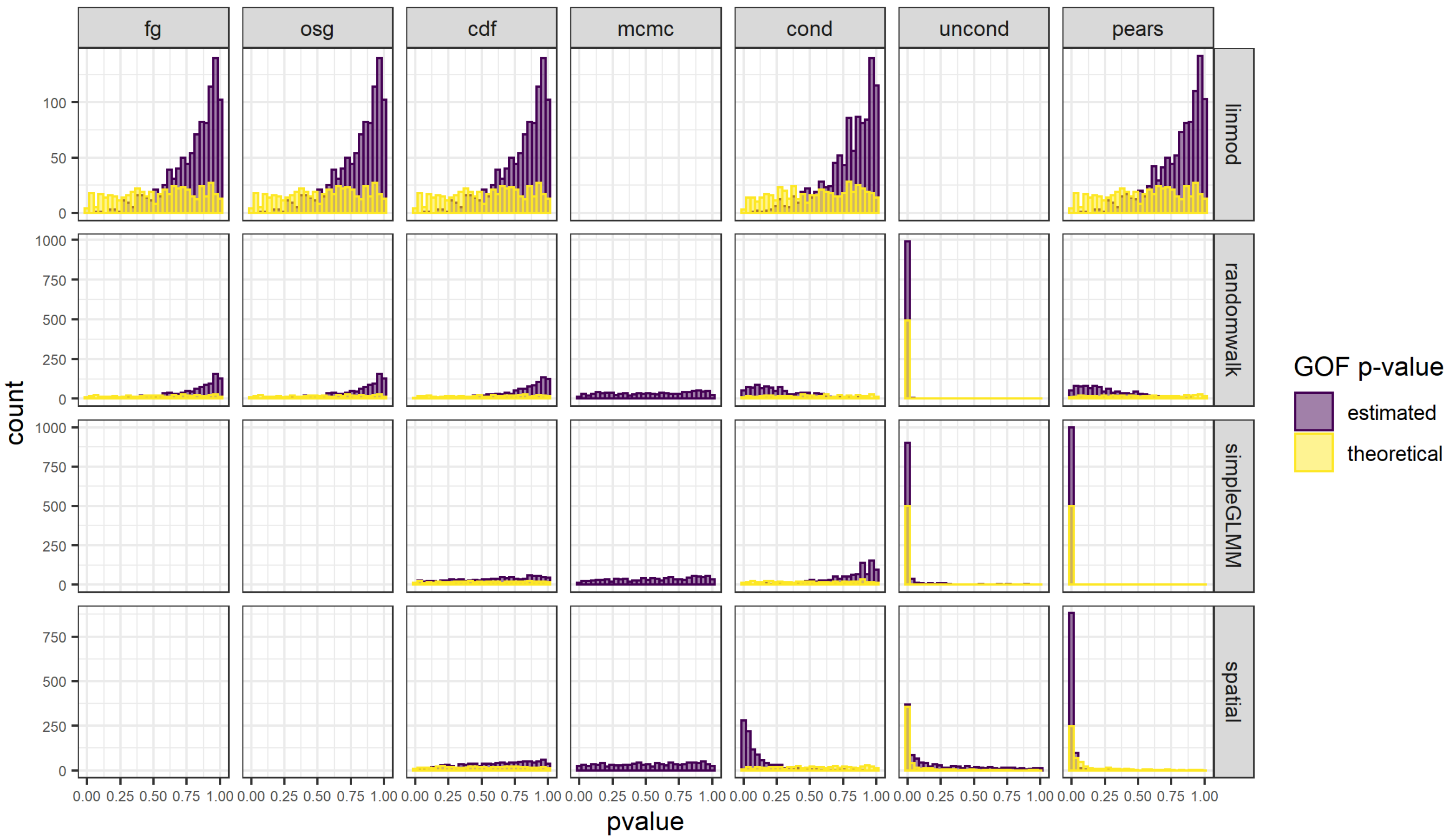
Model name	Model equations	Initial Values	Mis-specification
Linear model	$\eta = X\beta$ $y \sim N(\eta, \sigma)$	$X \sim N(0,1)$ $\beta = (4, -5)$ $\sigma = 1$	Lognormal error
Random walk	$\eta_i = \eta_{i-1} + \mu + \epsilon_i$ $\epsilon_i \sim N(0, \tau)$ $y \sim N(\eta, \sigma)$	$\mu = 0.75$ $\tau = 1$ $\sigma = 1$	Leave out drift term
GLMM	$u_j \sim N(0, \tau)$ $\eta_{i,j} = X_i\beta + u_j$ $y \sim Gamma(\frac{1}{\sigma^2}, \mu\sigma^2)$	$X \sim U(0,1)$ $\beta = (4, -.4)$ $\sigma^2 = 0.5$ $\tau^2 = 10$	Missing covariate
Spatial	$\omega \sim GMRF(Q[\kappa, \tau])$ $\eta_i = \beta_0 + \omega_i$ $y \sim Pois(\exp(\eta))$	$\kappa = \frac{\sqrt{8}}{50}$ $\sigma^2 = 2$ $\beta_0 = 0.5$	Lognormal spatial effect: $\eta_i = \beta_0 + \exp(\omega_i)$

Simulation Method

For each iteration (i=1000):

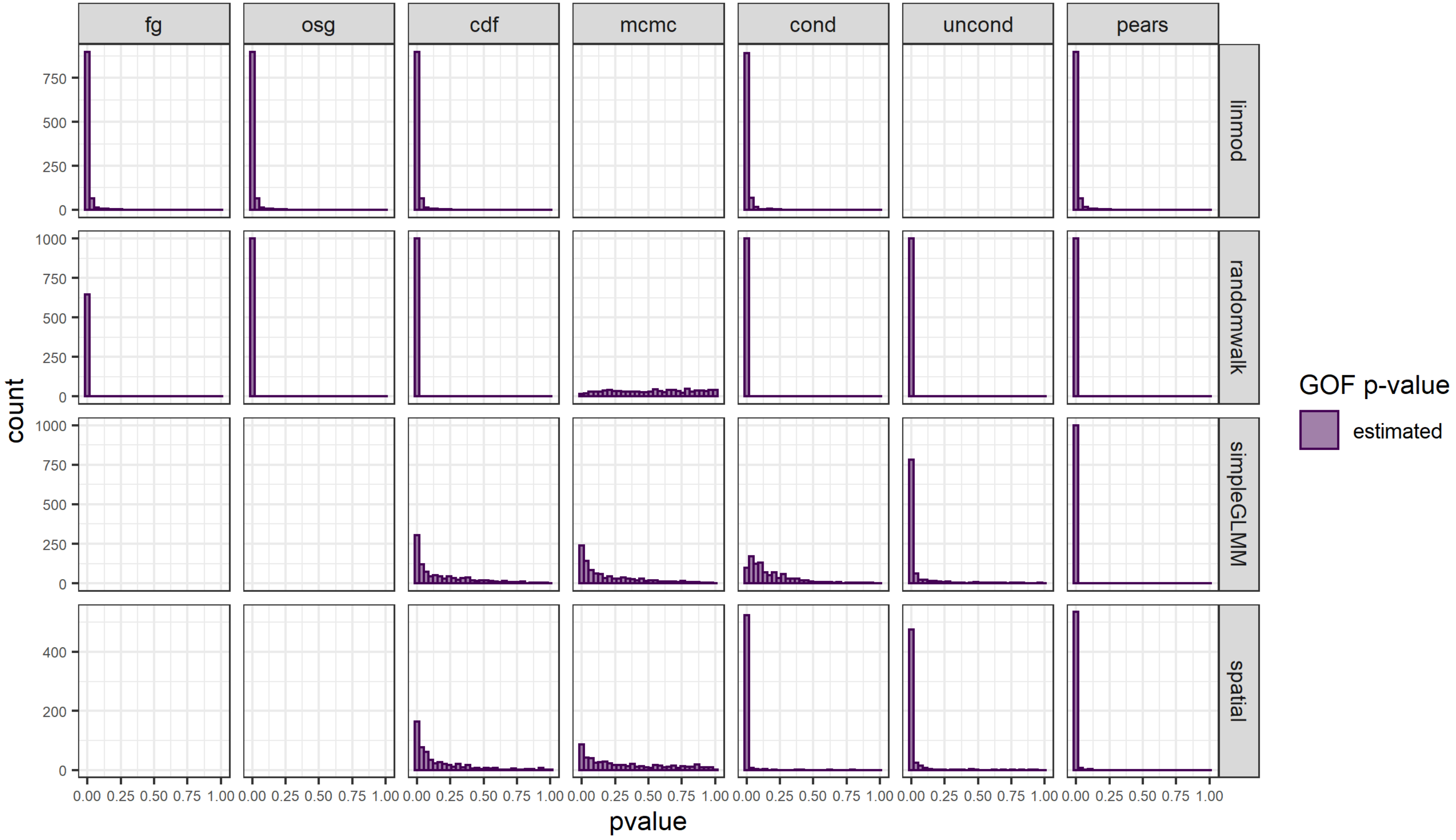
1. Data (n=100) were simulated for each model case.
2. Data were fit to the correctly specified (h0) and mis-specified (h1) models
3. OSA residuals were calculated for h0 and h1.
4. Simulation residuals were calculated for h0 and h1.
5. Steps 3 and 4 were repeated using true parameter values to calculate theoretical residuals.
6. GOF p-values were calculated using the KS test.

Correctly specified models



- Expectations:
  - Theoretical uniform
  - Estimated skewed towards 1
  - Pearson skewed towards 0
- Issues:
  - DHARMa unconditional
  - DHARMa conditional (spatial and randomwalk)

Mis-specified models

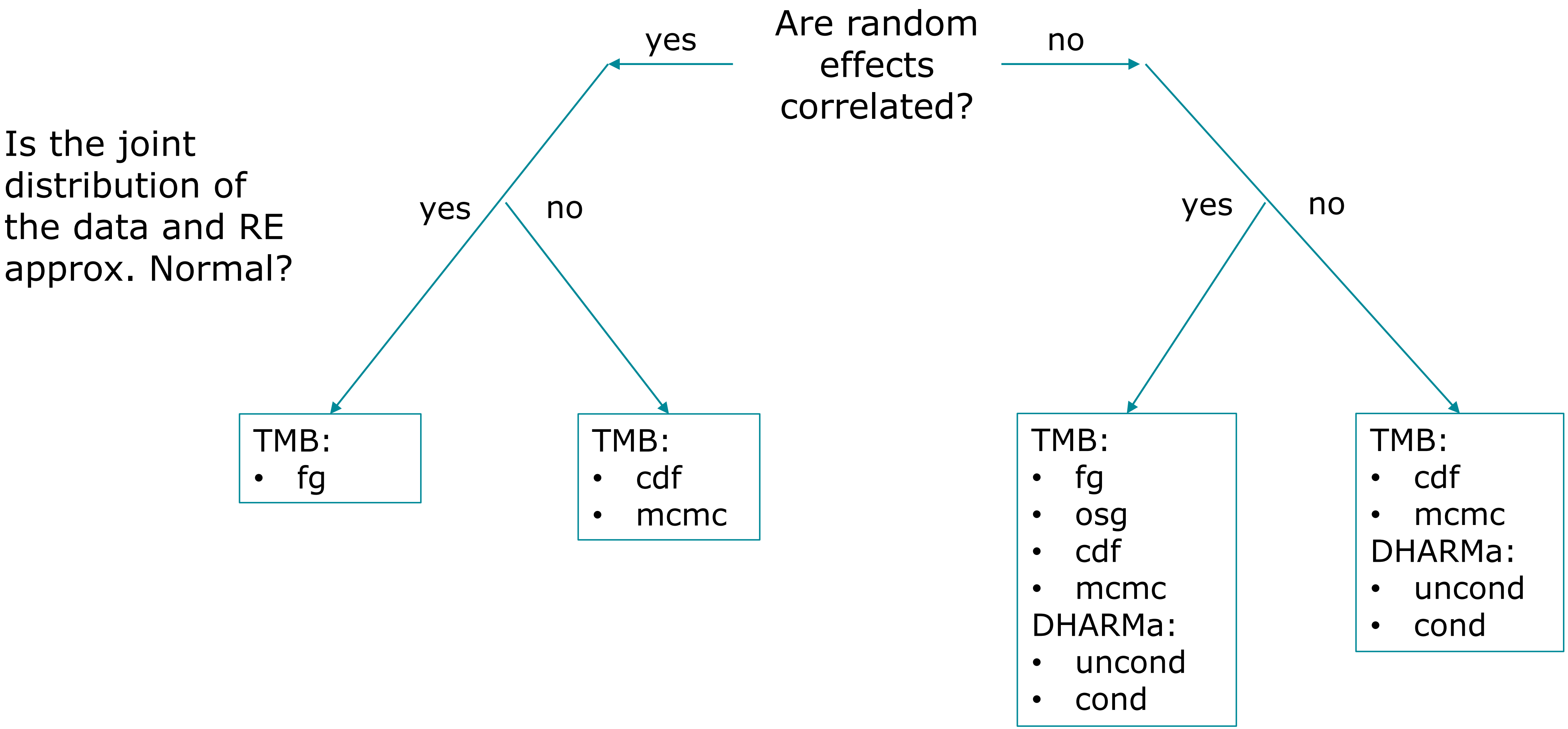


- Expectations:
  - Estimated skewed towards 0
- Issues:
  - randomwalk mcmc

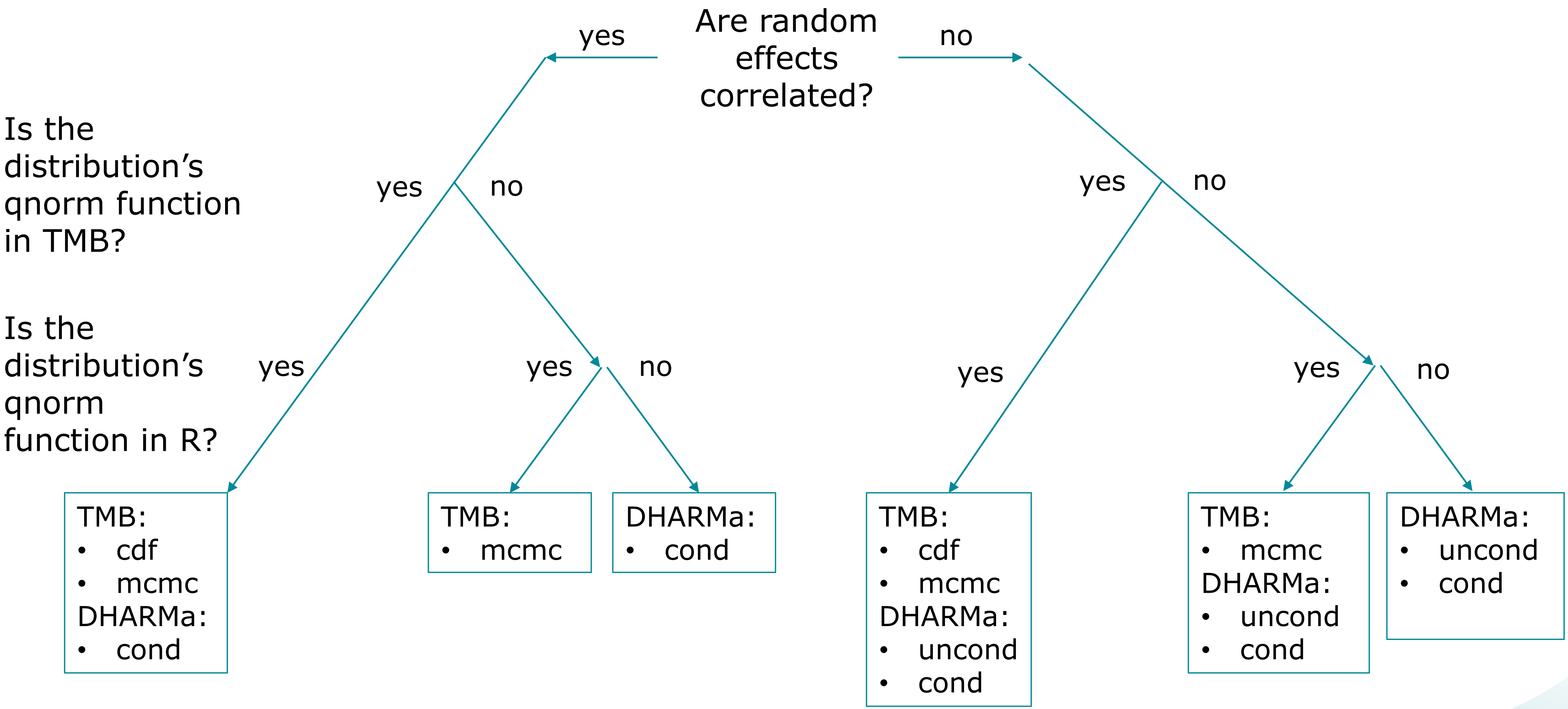
Simulation Summary

- Full Gaussian was fastest but also had the strictest assumptions
- The MCMC approach was also fast but tended to be less sensitive to mis-specification
- DHARMa methods were faster than one-step approaches but could not be used when random effects were MVN
- DHARMa methods are generalized so that they can be used when the cdf of the distribution is not well defined
- In general, random effects models are highly flexible and often lack the power needed to detect mis-specification

Decision Tree – Continuous Data



Decision Tree – Discrete or Hurdle Data



Next Steps

- Expand to include other validation tests
- Develop case studies
- Flush out guidelines
- How to validate integrated models?

Thank you!

References

- Dunn, PK & Smyth, GK. 1996. Randomized Quantile Residuals. Journal of Computational and Graphical Statistics. 236-244.
- Hartig, F. 2020. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level /Mixed) Regression Models. https://cran.r-project.org/package=DHARMa.
- Thygesen, U. et al., 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Sciences. 1-23.