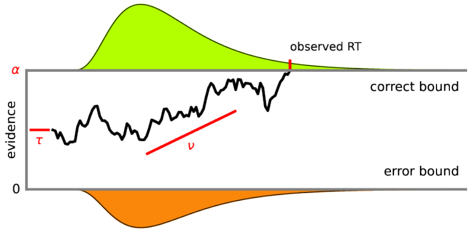
Robust, scalable cognitive psychometrics with the hierarchical Bayesian EZ-DDM

Adriana F. Chávez De la Peña[ab] and Joachim Vandekerckhove[a]

[a] University of California, Irvine
[b] Stanford University

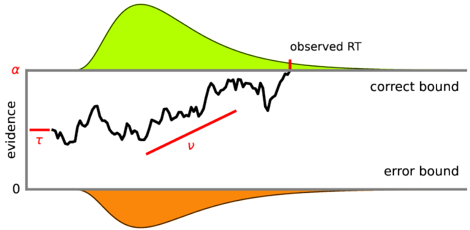
Cognitive psychometrics

- Cognitive models can double as measurement instruments (Batchelder, 2010; Lee and Wagenmakers, 2014)



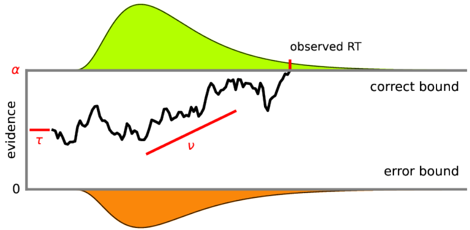
Cognitive psychometrics

- Cognitive models can double as measurement instruments (Batchelder, 2010; Lee and Wagenmakers, 2014)
- Choice and RT data turn into interpretable quantities: ability, caution, nondecision time (Ratcliff, 1978)



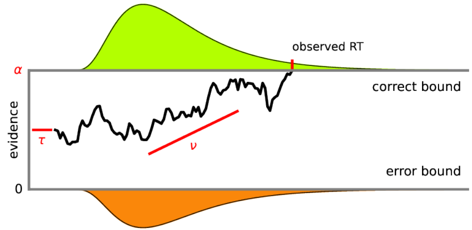
Cognitive psychometrics

- Cognitive models can double as measurement instruments (Batchelder, 2010; Lee and Wagenmakers, 2014)
- Choice and RT data turn into interpretable quantities: ability, caution, nondecision time (Ratcliff, 1978)
- Those quantities become our measurement targets



Cognitive psychometrics

- Cognitive models can double as measurement instruments (Batchelder, 2010; Lee and Wagenmakers, 2014)
- Choice and RT data turn into interpretable quantities: ability, caution, nondecision time (Ratcliff, 1978)
- Those quantities become our measurement targets
- The trick is doing this hierarchically, robustly, and at scale (Vandekerckhove et al., 2011)



"EZ" models and proxy likelihoods

- Often we don't need to fit every trial — we just want measurement

“EZ” models and proxy likelihoods

- Often we don’t need to fit every trial — we just want measurement
- So skip the raw data and build a likelihood for a few summary statistics instead

“EZ” models and proxy likelihoods

- Often we don’t need to fit every trial — we just want measurement
- So skip the raw data and build a likelihood for a few summary statistics instead
- That’s a proxy likelihood: summaries as a function of parameters

"EZ" models and proxy likelihoods

- Often we don't need to fit every trial — we just want measurement
- So skip the raw data and build a likelihood for a few summary statistics instead
- That's a proxy likelihood: summaries as a function of parameters
- EZ diffusion gives us closed-form equations (Wagenmakers et al., 2007):
  $(v, a, t_0) \ \rightarrow \ \left[P(\text{correct}),\ M_{RT},\ V_{RT}\right]$

"EZ" models and proxy likelihoods

- Often we don't need to fit every trial — we just want measurement
- So skip the raw data and build a likelihood for a few summary statistics instead
- That's a proxy likelihood: summaries as a function of parameters
- EZ diffusion gives us closed-form equations (Wagenmakers et al., 2007): $(v, a, t_0) \rightarrow \left[P(\text{correct}),\ M_{RT},\ V_{RT}\right]$
- Sampling theory then turns those summaries into a likelihood (Chávez De la Peña and Vandekerckhove, 2025):

$$C \sim \mathrm{Binomial}(N, R^{pr})$$

$$M^{\mathrm{obs}} \sim \mathcal{N}(M^{pr},\ V^{pr}/N)$$

$$V^{\mathrm{obs}} \sim \mathcal{N}(V^{pr},\ 2(V^{pr})^2/(N-1))$$

"EZ" models and proxy likelihoods

- Often we don't need to fit every trial — we just want measurement
- So skip the raw data and build a likelihood for a few summary statistics instead
- That's a proxy likelihood: summaries as a function of parameters
- EZ diffusion gives us closed-form equations (Wagenmakers et al., 2007):
  $(v, a, t_0) \;\rightarrow\; \left[P(\text{correct}),\; M_{RT},\; V_{RT}\right]$
- Sampling theory then turns those summaries into a likelihood (Chávez De la Peña and Vandekerckhove, 2025):

$$C \sim \mathrm{Binomial}(N, R^{pr})$$

$$M^{\mathrm{obs}} \sim \mathcal{N}(M^{pr}, V^{pr}/N)$$

$$V^{\mathrm{obs}} \sim \mathcal{N}(V^{pr}, 2(V^{pr})^2/(N-1))$$

- Drop that into JAGS (Plummer, 2003) and a full hierarchical fit takes seconds

Robust EZ-DDM

- The compression that makes EZ fast creates a weak point (Ratcliff, 2008)

Robust EZ-DDM

- The compression that makes EZ fast creates a weak point (Ratcliff, 2008)
- Mean RT and RT variance are easily thrown off by contaminants

Robust EZ-DDM

- The compression that makes EZ fast creates a weak point (Ratcliff, 2008)
- Mean RT and RT variance are easily thrown off by contaminants
- A few slow lapses per participant is enough to distort the fit

Robust EZ-DDM

- The compression that makes EZ fast creates a weak point (Ratcliff, 2008)
- Mean RT and RT variance are easily thrown off by contaminants
- A few slow lapses per participant is enough to distort the fit
- Robust alternative: keep the hierarchical proxy likelihood

Robust EZ-DDM

- The compression that makes EZ fast creates a weak point (Ratcliff, 2008)
- Mean RT and RT variance are easily thrown off by contaminants
- A few slow lapses per participant is enough to distort the fit
- Robust alternative: keep the hierarchical proxy likelihood
- Change only the observed RT summaries
  - $M_{RT} \to \text{Median}(RT)$
  - $V_{RT} \to \left(\frac{Q_3 - Q_1}{1.349}\right)^2$

Simulation stress test

- Four-way comparison:
  - Standard EZ $\times$ clean data
  - Robust EZ $\times$ clean data
  - Standard EZ $\times$ contaminated data
  - Robust EZ $\times$ contaminated data

Simulation stress test

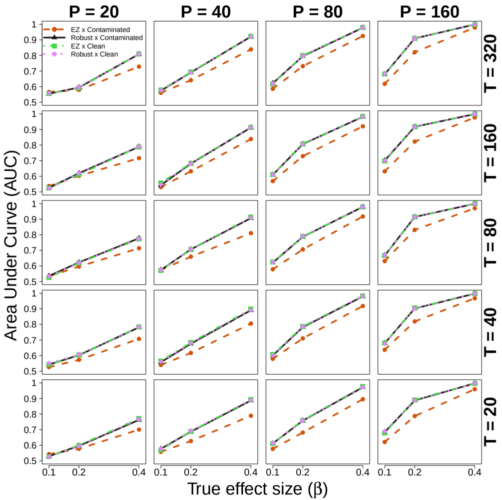
- Four-way comparison:
  - Standard EZ $\times$ clean data
  - Robust EZ $\times$ clean data
  - Standard EZ $\times$ contaminated data
  - Robust EZ $\times$ contaminated data
- Vary P, T, and true effect size β

Simulation stress test

- Four-way comparison:
  - Standard EZ $\times$ clean data
  - Robust EZ $\times$ clean data
  - Standard EZ $\times$ contaminated data
  - Robust EZ $\times$ contaminated data
- Vary P, T, and true effect size β
- Evaluate detection of condition effect on drift rate

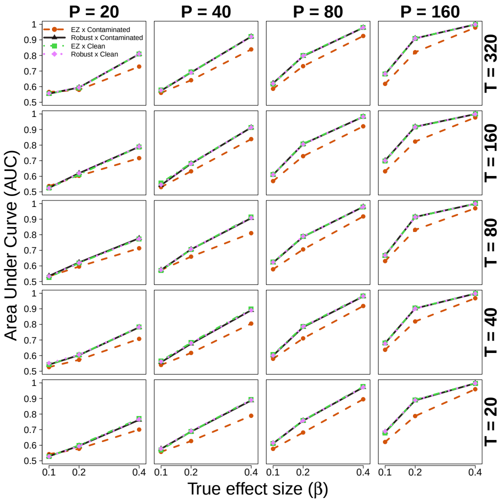
Simulation stress test

- Four-way comparison:
  - Standard EZ × clean data
  - Robust EZ × clean data
  - Standard EZ × contaminated data
  - Robust EZ × contaminated data
- Vary P, T, and true effect size β
- Evaluate detection of condition effect on drift rate
- AUC suffers only when contaminants are present and unaccounted for



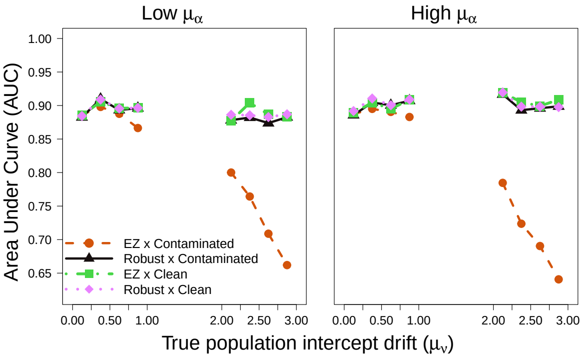
Simulation stress test

- Four-way comparison:
  - Standard EZ × clean data
  - Robust EZ × clean data
  - Standard EZ × contaminated data
  - Robust EZ × contaminated data
- Vary P, T, and true effect size β
- Evaluate detection of condition effect on drift rate
- AUC suffers only when contaminants are present and unaccounted for
- Robust EZ is otherwise as accurate as standard EZ



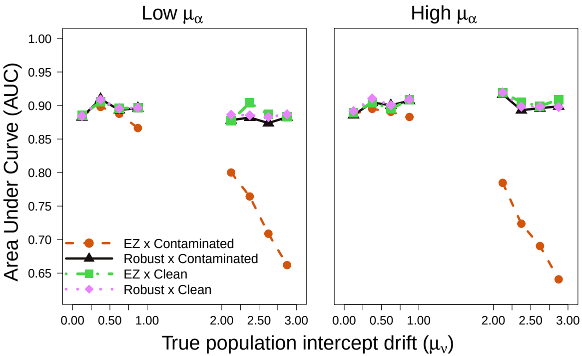
Why standard EZ fails

- The problem is strongest when drift is high



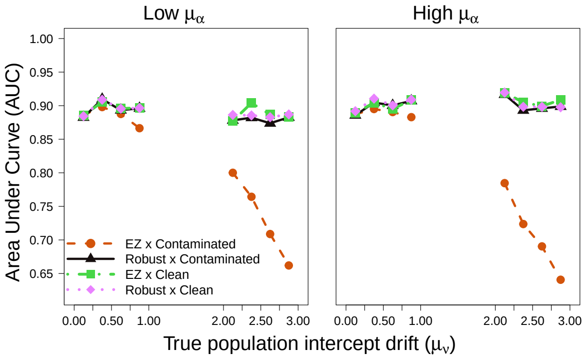
Why standard EZ fails

- The problem is strongest when drift is high
- In easy tasks, typical RTs are short, so slow contaminants dominate mean and variance



Why standard EZ fails

- The problem is strongest when drift is high
- In easy tasks, typical RTs are short, so slow contaminants dominate mean and variance
- The standard model attenuates μ_v and β; the robust model preserves the effect



Validating an example

A shape perception study

Validating an example

A shape perception study

Ask which stimulus-level properties move drift rate

Validating an example

A shape perception study

Ask which stimulus-level properties move drift rate

. . . and let any effects vary across participants and conditions

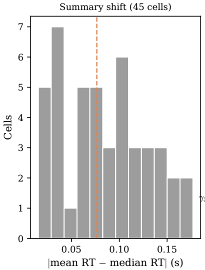
Validating an example

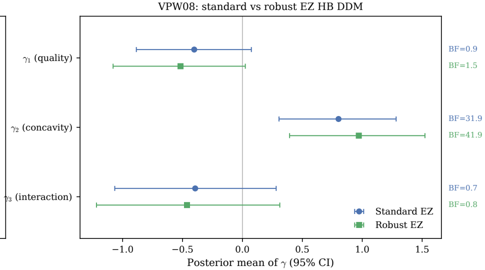
A shape perception study

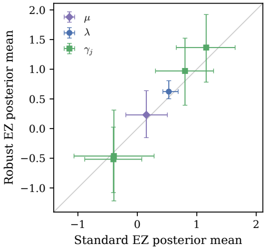
Ask which stimulus-level properties move drift rate

... and let any effects vary across participants and conditions

Results basically identical to standard EZDDM, despite RT differences of 100-150ms in half the cells







Second example

- Data from a large **lexical decision** task (*Is this a word? Y/N*)

Second example

- Data from a large **lexical decision** task (*Is this a word? Y/N*)
- Two modalities: Visual vs Auditory

Second example

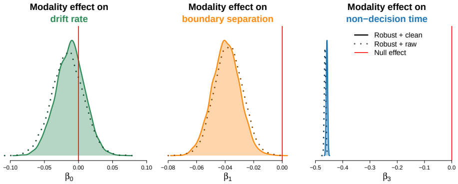
- Data from a large **lexical decision** task (*Is this a word? Y/N*)
- Two modalities: Visual vs Auditory
- 100 blocks of trials per participant

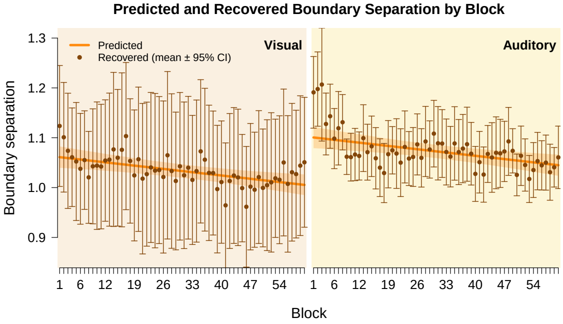
Second example

- Data from a large **lexical decision** task (*Is this a word? Y/N*)
- Two modalities: Visual vs Auditory
- 100 blocks of trials per participant

We inspect

- Modality effects



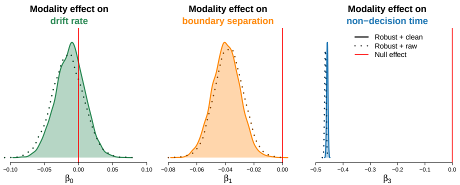


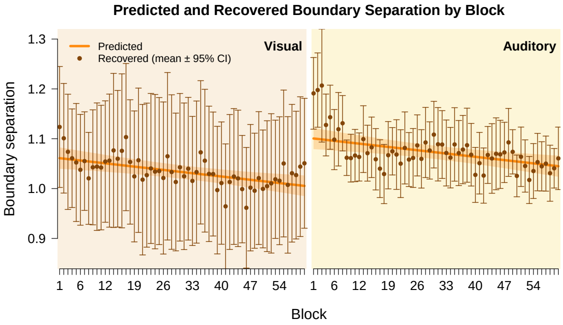
Second example

- Data from a large **lexical decision** task (*Is this a word? Y/N*)
- Two modalities: Visual vs Auditory
- 100 blocks of trials per participant

We inspect

- Modality effects
- Changes across blocks





But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

… but we still had to make a dicey independence assumption between the summaries

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

… but we still had to make a dicey independence assumption between the summaries

… and other interesting processes don't have that luxury – they're only simulatable

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

... but we still had to make a dicey independence assumption between the summaries

... and other interesting processes don't have that luxury – they're only simulatable

For example:

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

... but we still had to make a dicey independence assumption between the summaries

... and other interesting processes don't have that luxury – they're only simulatable

For example:

- collapsing bounds

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

... but we still had to make a dicey independence assumption between the summaries

... and other interesting processes don't have that luxury – they're only simulatable

For example:

- collapsing bounds
- time-varying parameters

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

... but we still had to make a dicey independence assumption between the summaries

... and other interesting processes don't have that luxury – they're only simulatable

For example:

- collapsing bounds
- time-varying parameters
- reinforcement-learning models of choice and RT

But EZ is the lucky case

EZ works out because Wagenmakers et al. (2007) solved the equations for us

... but we still had to make a dicey independence assumption between the summaries

... and other interesting processes don't have that luxury – they're only simulatable

For example:

- collapsing bounds
- time-varying parameters
- reinforcement-learning models of choice and RT

It would be great to if we could create proxy likelihoods without closed-form equations

Amortized synthetic likelihood

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

1. Simulate
draw $\theta \sim \pi(\theta)$; simulate summaries $s \mid \theta$; record mean $\bar{s}_\theta$ and covariance Σ_θ

2. Train emulator
dual-head net learns mean $\mu(\theta)$ and Cholesky $L(\theta)$; export as ONNX graph

3. Compile (JNNX)
ONNX → JAGS synthetic-likelihood module; new stochastic node {slug}_sl(θ, N)

4. Inference (JAGS)
obs_std $\sim$ {slug}_sl(θ, N); add priors, hierarchy, design — as usual

Amortized synthetic likelihood

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

$\theta \rightarrow$ simulate data

1. Simulate
draw $\theta \sim \pi(\theta)$; simulate summaries $s \mid \theta$; record mean $\bar{s}_\theta$ and covariance Σ_θ

2. Train emulator
dual-head net learns mean $\mu(\theta)$ and Cholesky $L(\theta)$; export as ONNX graph

3. Compile (JNNX)
ONNX $\rightarrow$ JAGS synthetic-likelihood module; new stochastic node {slug}_sl(θ, N)

4. Inference (JAGS)
obs_std $\sim$ {slug}_sl(θ, N); add priors, hierarchy, design — as usual

Amortized synthetic likelihood

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

$\theta \rightarrow$ simulate data

$\rightarrow$ compute summaries

1. Simulate
draw $\theta \sim \pi(\theta)$; simulate summaries $s \mid \theta$; record mean $\bar{s}_\theta$ and covariance Σ_θ

2. Train emulator
dual-head net learns mean $\mu(\theta)$ and Cholesky $L(\theta)$; export as ONNX graph

3. Compile (JNNX)
ONNX $\rightarrow$ JAGS synthetic-likelihood module; new stochastic node {slug}_sl(θ, N)

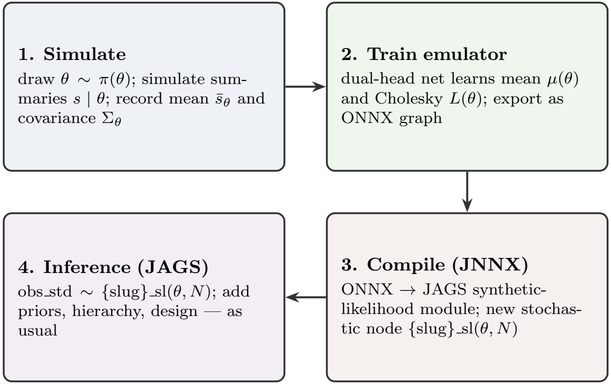
4. Inference (JAGS)
obs_std $\sim$ {slug}_sl(θ, N); add priors, hierarchy, design — as usual

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

$\theta \to$ simulate data

$\to$ compute summaries

$\to$ predict $\widehat{\mu}(\theta)$, $\widehat{\Sigma}(\theta)$



Amortized synthetic likelihood

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

$\theta \to$ simulate data

$\to$ compute summaries

$\to$ predict $\widehat{\mu}(\theta)$, $\widehat{\Sigma}(\theta)$

Then compile the trained network to a likelihood node in Bayesian graphical modeling software

1. Simulate
draw $\theta \sim \pi(\theta)$; simulate summaries $s \mid \theta$; record mean $\bar{s}_\theta$ and covariance Σ_θ

2. Train emulator
dual-head net learns mean $\mu(\theta)$ and Cholesky $L(\theta)$; export as ONNX graph

3. Compile (JNNX)
ONNX $\to$ JAGS synthetic-likelihood module; new stochastic node {slug}_sl(θ, N)

4. Inference (JAGS)
obs_std $\sim$ {slug}_sl(θ, N); add priors, hierarchy, design — as usual

Amortized synthetic likelihood

Amortization: Compute upfront, reuse many times (Fengler et al., 2021; Wood, 2010)

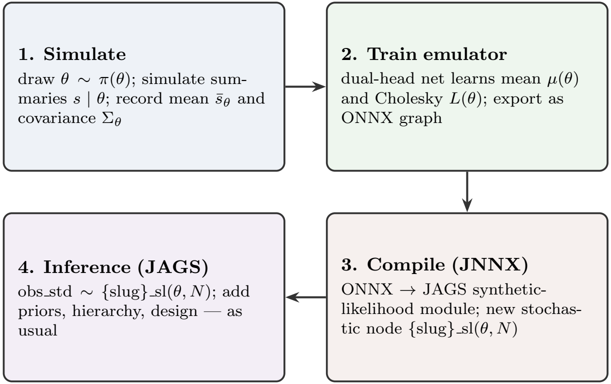
$\theta \rightarrow$ simulate data

$\rightarrow$ compute summaries

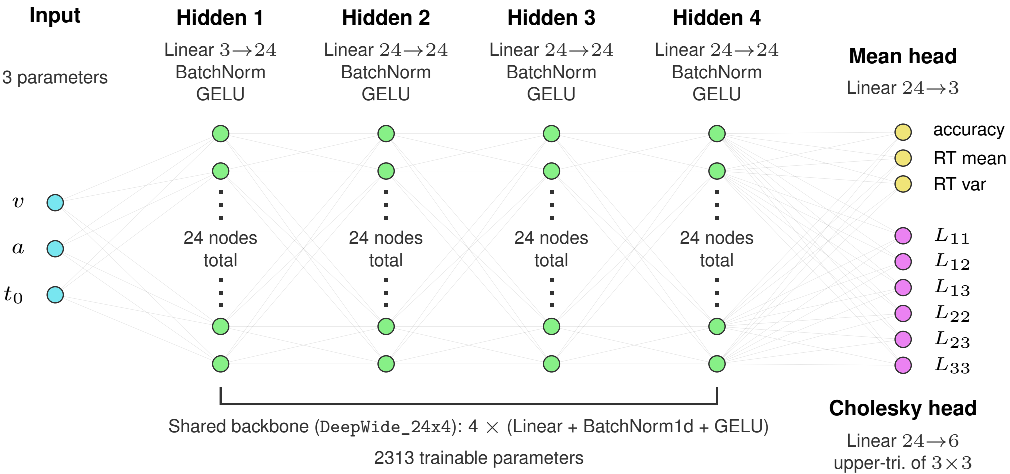
$\rightarrow$ predict $\widehat{\mu}(\theta)$, $\widehat{\Sigma}(\theta)$

Then compile the trained network to a likelihood node in Bayesian graphical modeling software

$\rightarrow$ rapidly iterate model structure



Emulator architecture



draw 20,000 θs from prior $\to$ 120 data sets of size 600 $\to$ compute $\widehat{\mu}(\theta)$ and $\widehat{\Sigma}(\theta)$

Example JAGS code

Compiled JAGS module includes a new stochastic node

$$(s_z \mid \theta, N) \sim \mathcal{N}\left(\widehat{\mu}\left(\theta\right), \Sigma_{\mathrm{samp}}\left(\theta, N\right) + \Sigma_{\mathrm{emu}}\right)$$

Example JAGS code

Compiled JAGS module includes a new stochastic node

$$(s_z \mid \theta, N) \sim \mathcal{N}\left(\widehat{\mu}\left(\theta\right), \Sigma_{\mathrm{samp}}\left(\theta, N\right) + \Sigma_{\mathrm{emu}}\right)$$

Works like any other node, takes parameters and N for summary statistics:

```
model {
  v  ~ dunif(-4.00, 4.00)
  a  ~ dunif( 0.50, 4.00)
  t0 ~ dunif( 0.05, 0.65)
  obs_std[1:3] ~ ddm3mv_sl(v, a, t0, n_trials)
}
```

Example JAGS code

Compiled JAGS module includes a new stochastic node

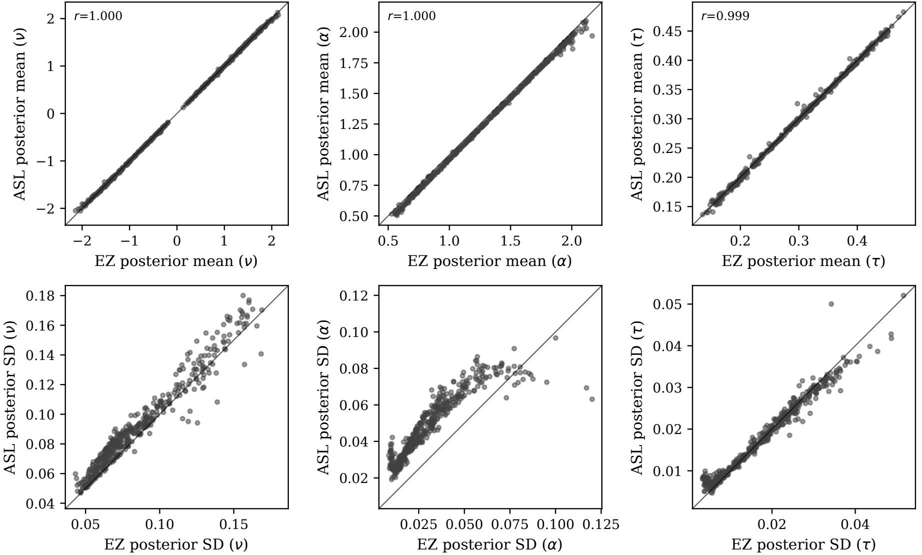
$$(s_z \mid \theta, N) \sim \mathcal{N}\left(\widehat{\mu}\left(\theta\right), \Sigma_{\mathrm{samp}}\left(\theta, N\right) + \Sigma_{\mathrm{emu}}\right)$$

Works like any other node, takes parameters and N for summary statistics:

```
model {
  v  ~ dunif(-4.00, 4.00)
  a  ~ dunif( 0.50, 4.00)
  t0 ~ dunif( 0.05, 0.65)
  obs_std[1:3] ~ ddm3mv_sl(v, a, t0, n_trials)
}
```

5000 MCMC steps in < 30s.

Validation against closed-form EZ DDM



Validation ladder

1. **DDM3** recovers the analytic EZ case

Validation ladder

1. **DDM3** recovers the analytic EZ case

$\rightarrow$ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

Validation ladder

1. **DDM3** recovers the analytic EZ case

$\rightarrow$ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

2. **DDM4** adds starting bias

Validation ladder

1. **DDM3** recovers the analytic EZ case

$\rightarrow$ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

2. **DDM4** adds starting bias

$\rightarrow$ uses response-specific summaries

Validation ladder

1. **DDM3** recovers the analytic EZ case

→ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

2. **DDM4** adds starting bias

→ uses response-specific summaries

→ recovery and calibration are excellent

Validation ladder

1. **DDM3** recovers the analytic EZ case

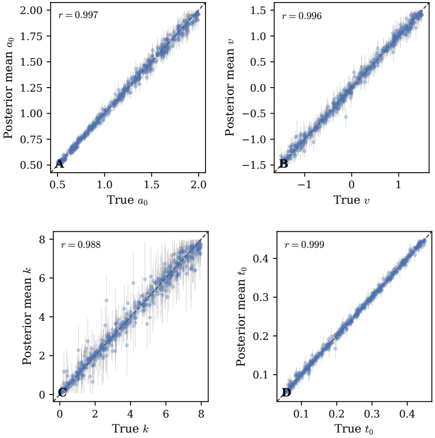
→ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

2. **DDM4** adds starting bias

→ uses response-specific summaries

→ recovery and calibration are excellent

3. **DDM-COLLAPSE** bounds collapse according to $b(t) = a_0 \Big/ \left(1 + e^{kt}\right)$



Validation ladder

1. **DDM3** recovers the analytic EZ case

→ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

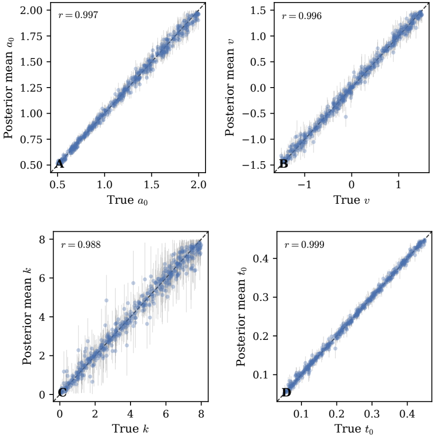
2. **DDM4** adds starting bias

→ uses response-specific summaries

→ recovery and calibration are excellent

3. **DDM-COLLAPSE** bounds collapse according to $b(t) = a_0 \Big/ \left(1 + e^{kt}\right)$

→ uses quantiles as summaries



Validation ladder

1. **DDM3** recovers the analytic EZ case

$\to$ with better calibration of $\sigma(\alpha)$ due to multivariate modeling of summary vector

2. **DDM4** adds starting bias

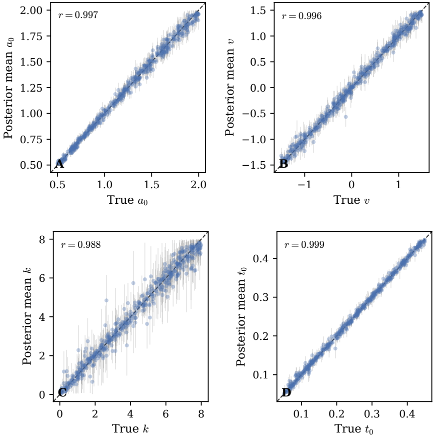
$\to$ uses response-specific summaries

$\to$ recovery and calibration are excellent

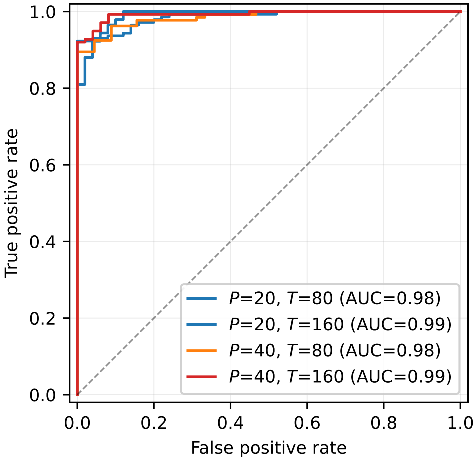
3. **DDM-COLLAPSE** bounds collapse according to $b(t) = a_0 \Big/ \left(1 + e^{kt}\right)$

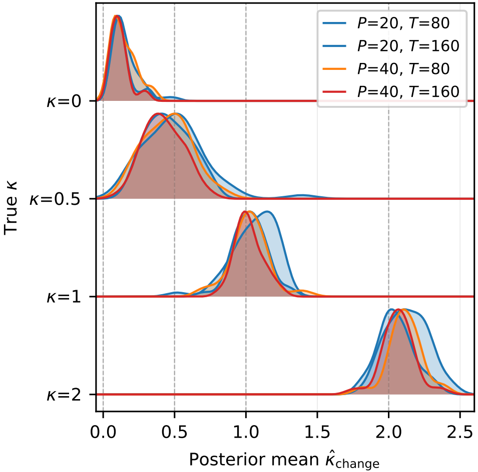
$\to$ uses quantiles as summaries

$\to$ recovery and calibration are excellent



Hypothesis test calibration





Summary

- **Robust EZ-DDM**: proxy likelihoods that hold up on RT data with contaminants

Summary

- **Robust EZ-DDM**: proxy likelihoods that hold up on RT data with contaminants
- **Amortized synthetic likelihood**: the same proxy trick works for models we can only simulate

Summary

- **Robust EZ-DDM**: proxy likelihoods that hold up on RT data with contaminants
- **Amortized synthetic likelihood**: the same proxy trick works for models we can only simulate
- Adds to our cognitive psychometric toolbox: turns cognitive process models into measurement tools that are:
  - fast enough to use regularly
  - reusable to iterate over structures
  - flexible with regard to process model assumptions

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In Embretson, S. E., editor, *Measuring psychological constructs: Advances in model-based approaches*, pages 71–93. American Psychological Association, Washington, DC.

Chávez De la Peña, A. F. and Vandekerckhove, J. (2025). An EZ Bayesian hierarchical drift diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 32:3067–3087.

Fengler, A., Govindarajan, L. N., Chen, T., and Frank, M. J. (2021). Likelihood approximation networks (lans) for fast inference of simulation models in cognitive neuroscience. *eLife*, 10:e65074.

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85:59–108.

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin & Review*, 15(6):1218–1228.

Vandekerckhove, J., Tuerlinckx, F., and Lee, M. D. (2011). Hierarchical diffusion models for two–choice response times. *Psychological Methods*, 16:44–62.

Wagenmakers, E.-J., van der Maas, H. J. L., and Grasman, R. P. P. P. (2007). An EZ–diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14:3–22.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.